

Ultra-high Dimensional Multiple Output Learning With Simultaneous Orthogonal Matching Pursuit: A Sure Screening Approach

Mladen Kolar and Eric P. Xing
School of Computer Science
Carnegie Mellon University

December 20, 2010

Abstract

We propose a novel application of the Simultaneous Orthogonal Matching Pursuit (S-OMP) procedure for sparsistent variable selection in ultra-high dimensional multi-task regression problems. Screening of variables, as introduced in Fan and Lv (2008), is an efficient and highly scalable way to remove many irrelevant variables from the set of all variables, while retaining all the relevant variables. S-OMP can be applied to problems with hundreds of thousands of variables and once the number of variables is reduced to a manageable size, a more computationally demanding procedure can be used to identify the relevant variables for each of the regression outputs. To our knowledge, this is the first attempt to utilize relatedness of multiple outputs to perform fast screening of relevant variables. As our main theoretical contribution, we prove that, asymptotically, S-OMP is guaranteed to reduce an ultra-high number of variables to below the sample size without losing true relevant variables. We also provide formal evidence that a modified Bayesian information criterion (BIC) can be used to efficiently determine the number of iterations in S-OMP. We further provide empirical evidence on the benefit of variable selection using multiple regression outputs jointly, as opposed to performing variable selection for each output separately. The finite sample performance of S-OMP is demonstrated on extensive simulation studies, and on a genetic association mapping problem.

Keywords: Adaptive Lasso; Greedy forward regression; Orthogonal matching pursuit; Multi-output regression; Multi-task learning; Simultaneous orthogonal matching pursuit; Sure screening; Variable selection

1 Introduction

Multiple output regression, also known as multi-task regression, with *ultra-high dimensional* inputs commonly arise in problems such as genome-wide association (GWA) mapping in genetics, or stock portfolio prediction in finance. For

example, in a GWA mapping problem, the goal is to find a small set of relevant single-nucleotide polymorphisms (SNP) (*covariates, or inputs*) that account for variations of a large number of gene expression or clinical traits (*responses, or outputs*), through a response function that is often modeled via a regression. However, this is a very challenging problem for current statistical methods since the number of input variables is likely to reach millions, prohibiting even usage of scalable implementation of Lasso-like procedures for model selection, which are a convex relaxation of a combinatorial subset selection search. Furthermore, the outputs in a typical multi-task regression problem are not independent of each other, therefore the discovery of truly relevant inputs has to take into consideration of potential joint effects induced by coupled responses. To appreciate this better, consider again the GWA example. Typically, genes in a biological pathway are co-expressed as a module and it is often assumed that a causal SNP affects multiple genes in one pathway, but not all of the genes in the pathway. In order to effectively reduce the dimensionality of the problem and to detect the causal SNPs, it is very important to look at how SNPs affect all genes in a biological pathway. Since the experimentally collected data is usually very noisy, regressing genes individually onto SNPs may not be sufficient to identify the relevant SNPs that are only weakly marginally correlated with each individual gene in a module. However, once the whole biological pathway is examined, it is much easier to find such causal SNPs. In this paper, we demonstrate that the Simultaneous Orthogonal Matching Pursuit (S-OMP) (Tropp et al., 2006) can be used to quickly reduce the dimensionality of such problems, without losing any of the relevant variables.

From a computational point of view, as the dimensionality of the problem and the number of outputs increase, it can become intractable to solve the underlying convex programs commonly used to identify relevant variables in multi-task regression problems. Previous work by Liu et al. (2009), Lounici et al. (2009) and Kim et al. (2009), for example, do not scale well to settings when the number of variables exceeds $\gtrsim 10000$ and the number of outputs exceeds $\gtrsim 1000$, as in typical genome-wide association studies. Furthermore, since the estimation error of the regression coefficients depends on the number of variables in the problem, variable selection can improve convergence rates of estimation procedures. These concerns motivate us to propose and study the S-OMP as a fast way to remove irrelevant variables from an ultra-high dimensional space.

Formally, the GWA mapping problem, which we will use as an illustrative example both in here for model formulation and later for empirical experimental validation, can be cast as a variable selection problem in a multiple output regression model:

$$\mathbf{Y} = \mathbf{X}\mathbf{B} + \mathbf{W} \quad (1)$$

where $\mathbf{Y} = [\mathbf{y}_1, \dots, \mathbf{y}_T] \in \mathbb{R}^{n \times T}$ is a matrix of outputs, whose column \mathbf{y}_t is an n -vector for the t -th output (i.e., gene), $\mathbf{X} \in \mathbb{R}^{n \times p}$ is a random design matrix, of which each row \mathbf{x}_i denotes a p -dimensional input, $\mathbf{B} = [\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_T] \in \mathbb{R}^{p \times T}$ is the matrix of regression coefficients and $\mathbf{W} = [\boldsymbol{\epsilon}_1, \dots, \boldsymbol{\epsilon}_T] \in \mathbb{R}^{n \times T}$ is a matrix of IID random noise, independent of \mathbf{X} . Throughout the paper we are going to

assume that the columns of \mathbf{B} are jointly sparse, as we precisely specify below. Note that if different columns of \mathbf{B} do not share any underlying structure, the model in (1) can be estimated by fitting each of the tasks separately.

We are interested in estimating the regression coefficients, under the assumption that they share a common structure, e.g., there exist a subset of variables with non-zero coefficients for more than one regression output. We informally refer to such outputs as related. Such a variable selection problem can be formalized in two ways: (1) the *union support* recovery of \mathbf{B} , as defined in Obozinski et al. (2010), where a subset of variables is selected that affect at least one output; (2) the *exact support* recovery of \mathbf{B} , where the exact positions of non-zero elements in \mathbf{B} are estimated. In this paper, we concern ourselves with exact support recovery, which is of particular importance in problems like GWA mapping (Kim and Xing, 2009) or biological network estimation (Peng et al., 2008). Under such a multi-task setting, two interesting questions naturally follow: i) how can information be shared between related outputs in order to improve the predictive accuracy and the rate of convergence of the estimated regression coefficients over the independent estimation on each output separately; ii) how to select relevant variables more accurately based on information from related outputs. To address these two questions, one line of research (e.g., Zhang, 2006; Liu et al., 2009; Lounici et al., 2009) has looked into the following estimation procedure leveraging a *multi-task regularization*:

$$\hat{\mathbf{B}} = \underset{\boldsymbol{\beta}_t \in \mathbb{R}^p, t \in [T]}{\operatorname{argmin}} \sum_{t=1}^T \|\mathbf{y}_t - \mathbf{X}\boldsymbol{\beta}_t\|_2^2 + \lambda \sum_{j=1}^p \operatorname{pen}(\beta_{1,j}, \dots, \beta_{T,j}), \quad (2)$$

with $\operatorname{pen}(a_1, \dots, a_T) = \max_{t \in [T]} |a_t|$ or $\operatorname{pen}(a_1, \dots, a_T) = \sqrt{\sum_{t \in [T]} a_t^2}$ for a vector $\mathbf{a} \in \mathbb{R}^T$. Under an appropriate choice of the penalty parameter λ , the estimator $\hat{\mathbf{B}}$ has many rows equal to zero, which correspond to irrelevant variables. However, solving (2) can be computationally prohibitive.

In this paper, we consider an ultra-high dimensional setting for the aforementioned multi-task regression problem, where the number of variables p is much higher than the sample size n , e.g. $p = \mathcal{O}(\exp(n^{\delta_p}))$ for a positive constant δ_p , but the regression coefficients $\boldsymbol{\beta}_t$ are sparse, i.e., for each task t , there exist a very small number of variables that are relevant to the output. Under the sparsity assumption, it is highly important to efficiently select the relevant variables in order to improve the accuracy of the estimation and prediction, and to facilitate the understanding of the underlying phenomenon for domain experts. In the seminal paper of Fan and Lv (2008), the concept of *sure screening* was introduced, which leads to a sequential variable selection procedure that keeps all the relevant variables with high probability in ultra-high dimensional *uni-output regression*. In this paper, we propose the S-OMP procedure, which enjoys *sure screening* property in ultra-high dimensional *multiple output regression* as defined in (1). To perform *exact support* recovery, we further propose a two-step procedure that first use S-OMP to screen the variables, i.e., select a subset of variables that contain all the true variables; and then use the adap-

tive Lasso (ALasso) (Zou, 2006) to further select a subset of screened variables for each task. We show, both theoretically and empirically, that our procedure ensure sparsistent recovery of relevant variables. To the best of our knowledge, this is the first attempt to analyze the sure screening property in the ultra-high dimensional space using the shared information from the multiple regression outputs.

1.1 Related Work

The model given in (1) has been used in many different domains ranging from multivariate regression (Obozinski et al., 2009; Negahban and Wainwright, 2009) and sparse approximation (Tropp et al., 2006) to neural science (Liu et al., 2009), multi-task learning (Lounici et al., 2009; Argyriou et al., 2008) and biological network estimation (Peng et al., 2008). A number of authors has provided theoretical understanding of the estimation in the model using the convex program (2) to estimate $\hat{\mathbf{B}}$. Lounici et al. (2009) showed the benefits of the joint estimation, when there is a small set of variables common to all outputs and the number of outputs is large. Obozinski et al. (2009) and Negahban and Wainwright (2009) analyzed the consistent recovery of the union support. Negahban and Wainwright (2009) provided the analysis of the exact support recovery for a special case with two outputs.

The Orthogonal Matching Pursuit (OMP) has been analyzed before in the literature (see, e.g., Zhang, 2009; Lozano et al., 2009; Wang, 2009; Barron et al., 2008). In particular, our work should be contrasted to Wang (2009), which showed that the OMP has the sure screening property in a linear regression with a single output, and to the exact variable selection property of the OMP analyzed in Zhang (2009) and Lozano et al. (2009). The exact variable selection requires much stronger assumptions on the design, such as the irrepresentable condition, that are hard to satisfy in the ultra-high dimensional setting. On the other hand, the sure screening property can be shown to hold under much weaker assumptions.

In this paper, we make the following novel contributions: i) we prove that the S-OMP can be used for the ultra-high dimensional variable screening in multiple output regression problems and demonstrate its performance on extensive numerical studies; ii) we show that a two step procedure can be used to select exactly the relevant variables for each task; and iii) we prove that a modification of the BIC score (Chen and Chen, 2008) can be used to select the number of steps in the S-OMP.

The rest of the article is organized as follows. In Section 2, we introduce the simultaneous greedy forward regression and propose our approach to the exact support estimation. Theoretical guarantees of the methods are given in Section 3. Section 4 is devoted to extensive numerical simulations. An application to the real world problem in association mapping is demonstrated in Section 5. We conclude with discussion in Section 6. Proofs are deferred to Appendix.

2 Methodology

2.1 The model and notation

We will consider a slightly more general model

$$\begin{aligned} \mathbf{y}_1 &= \mathbf{X}_1 \boldsymbol{\beta}_1 + \boldsymbol{\epsilon}_1 \\ \mathbf{y}_2 &= \mathbf{X}_2 \boldsymbol{\beta}_2 + \boldsymbol{\epsilon}_2 \\ &\dots \\ \mathbf{y}_T &= \mathbf{X}_T \boldsymbol{\beta}_T + \boldsymbol{\epsilon}_T, \end{aligned} \tag{3}$$

than the one given in (1). The model in (1) is a special case of the model in (3), with all the design matrices $\{\mathbf{X}_t\}_{t \in [T]}$ being equal and $[T]$ denoting the set $\{1, \dots, T\}$. Assume that for all $t \in [T]$, $\mathbf{X}_t \in \mathbb{R}^{n \times p}$. For the design \mathbf{X}_t , we denote $\mathbf{X}_{t,j}$ the j -th column (i.e., dimension), $\mathbf{x}_{t,i}$ the i -th row (i.e., instance) and $x_{t,ij}$ the element at (i, j) . Denote $\boldsymbol{\Sigma}_t = \text{Cov}(\mathbf{x}_{t,i})$. Without loss of generality, we assume that $\text{Var}(y_{t,i}) = 1$, $\mathbb{E}(x_{t,ij}) = 0$ and $\text{Var}(x_{t,ij}) = 1$. The noise $\boldsymbol{\epsilon}_t$ is zero mean and $\text{Cov}(\boldsymbol{\epsilon}_t) = \sigma^2 \mathbf{I}_{n \times n}$. We assume that the number of variables $p \gg n$ and that the vector of regression coefficients $\boldsymbol{\beta}_t$ are jointly sparse, that is, there exist a small number of variables that are relevant for the most of the T regression problems. Put another way, the matrix $B = [\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_T]$ has only a small number of non-zero rows. Let $\mathcal{M}_{*,t}$ denote the set of non-zero coefficients of $\boldsymbol{\beta}_t$ and $\mathcal{M}_* = \cup_{t=1}^T \mathcal{M}_{*,t}$ denote the set of all relevant variables, i.e., variables with non-zero coefficient in at least one of the tasks. For an arbitrary set $\mathcal{M} \subseteq \{1, \dots, p\}$, $\mathbf{X}_{t,\mathcal{M}}$ denotes the design with columns indexed by \mathcal{M} , $\mathbf{B}_{\mathcal{M}}$ denotes the rows of \mathbf{B} indexed by \mathcal{M} and $\mathbf{B}_j = (\beta_{1,j}, \dots, \beta_{T,j})'$. The cardinality of the set \mathcal{M} is denoted as $|\mathcal{M}|$. Let $s := |\mathcal{M}_*|$ denote the total number of relevant variables, so under the sparsity assumption we have $s < n$. For a square matrix \mathbf{A} , $\Lambda_{\min}(\mathbf{A})$ and $\Lambda_{\max}(\mathbf{A})$ are used to denote the minimum and the maximum eigenvalue, respectively. For a different matrix $\mathbf{A} = [a_{ij}] \in \mathbb{R}^{p \times T}$, we define $\|\mathbf{A}\|_{2,1} := \sum_{i \in [p]} \sqrt{\sum_{j \in [T]} a_{ij}^2}$. Lastly, we use $[p]$ to denote the set $\{1, \dots, p\}$.

2.2 Simultaneous Orthogonal Matching Pursuit

We propose a Simultaneous Orthogonal Matching Pursuit procedure for ultra high-dimensional variable selection in the multi-task regression problem, which is outlined in Algorithm 1. Before describing the algorithm, we introduce some additional notation. For an arbitrary subset $\mathcal{M} \subseteq [p]$ of variables, let $\mathbf{H}_{t,\mathcal{M}}$ be the orthogonal projection matrix onto $\text{Span}(\mathbf{X}_{t,\mathcal{M}})$, i.e.,

$$\mathbf{H}_{t,\mathcal{M}} = \mathbf{X}_{t,\mathcal{M}} (\mathbf{X}_{t,\mathcal{M}}' \mathbf{X}_{t,\mathcal{M}})^{-1} \mathbf{X}_{t,\mathcal{M}}', \tag{4}$$

and define the residual sum of squares (RSS) as

$$\text{RSS}(\mathcal{M}) = \sum_{t=1}^T \mathbf{y}_t' (\mathbf{I}_{n \times n} - \mathbf{H}_{\mathcal{M}}) \mathbf{y}_t. \tag{5}$$

Algorithm 1 Group Forward Regression

Input: Dataset $\{\mathbf{X}_t, \mathbf{y}_t\}_{t=1}^T$ **Output:** Sequence of selected models $\{\mathcal{M}^{(k)}\}_{k=0}^{n-1}$

```
1: Set  $\mathcal{M}^{(0)} = \emptyset$ 
2: for  $k = 1$  to  $n - 1$  do
3:   for  $j = 1$  to  $p$  do
4:      $\tilde{\mathcal{M}}_j^{(k)} = \mathcal{M}^{(k-1)} \cup \{j\}$ 
5:      $\mathbf{H}_{t,j} = \mathbf{X}_{t,\tilde{\mathcal{M}}_j^{(k)}} (\mathbf{X}_{t,\tilde{\mathcal{M}}_j^{(k)}}' \mathbf{X}_{t,\tilde{\mathcal{M}}_j^{(k)}})^{-1} \mathbf{X}_{t,\tilde{\mathcal{M}}_j^{(k)}}'$ 
6:      $\text{RSS}(\tilde{\mathcal{M}}_j^{(k)}) = \sum_{t=1}^T \mathbf{y}_t' (\mathbf{I}_{n \times n} - \mathbf{H}_{t,j}) \mathbf{y}_t$ 
7:   end for
8:    $\hat{f}_k = \text{argmin}_{j \in \{1, \dots, p\} \setminus \mathcal{M}^{(k-1)}} \text{RSS}(\tilde{\mathcal{M}}_j^{(k)})$ 
9:    $\mathcal{M}^{(k)} = \mathcal{M}^{(k-1)} \cup \{\hat{f}_k\}$ 
10: end for
```

The algorithm starts with an empty set $\mathcal{M}^{(0)} = \emptyset$. We recursively define the set $\mathcal{M}^{(k)}$ based on the set $\mathcal{M}^{(k-1)}$. The set $\mathcal{M}^{(k)}$ is obtained by adding a variable indexed by $\hat{f}_k \in [p]$, which minimizes $\text{RSS}(\mathcal{M}^{(k-1)} \cup j)$ over the set $[p] \setminus \mathcal{M}^{(k-1)}$, to the set $\mathcal{M}^{(k-1)}$. Repeating the algorithm for $n - 1$ steps, a sequence of nested sets $\{\mathcal{M}^{(k)}\}_{k=0}^{n-1}$ is obtained, with $\mathcal{M}^{(k)} = \{\hat{f}_1, \dots, \hat{f}_k\}$.

To practically select one of the sets of variables from $\{\mathcal{M}^{(k)}\}_{k=0}^{n-1}$, we minimize the modified BIC criterion (Chen and Chen, 2008), which is defined as

$$\text{BIC}(\mathcal{M}) = \log \left(\frac{\text{RSS}(\mathcal{M})}{nT} \right) + \frac{|\mathcal{M}|(\log(n) + 2 \log(p))}{n} \quad (6)$$

with $|\mathcal{M}|$ denoting the number of elements of the set \mathcal{M} . Let

$$\hat{s} = \text{argmin}_{k \in \{0, \dots, n-1\}} \text{BIC}(\mathcal{M}^{(k)}),$$

so that the selected model is $\mathcal{M}^{(\hat{s})}$.

Remark: The S-OMP algorithm is outlined only conceptually in this section. The steps 5 and 6 of the algorithm can be implemented efficiently using the progressive Cholesky decomposition see, e.g., Cotter et al. (1999). A computationally costly step 5 involves inversion of the matrix $\mathbf{X}_{t,\mathcal{M}}' \mathbf{X}_{t,\mathcal{M}}$, however, it can be seen from the algorithm that the matrix $\mathbf{X}_{t,\mathcal{M}}' \mathbf{X}_{t,\mathcal{M}}$ is updated in each iteration by simply appending a row and a column to it. Therefore, its Cholesky factorization can be efficiently computed based on calculation involving only the last row. A detailed implementation of the orthogonal matching pursuit algorithm based on the progressive Cholesky decomposition can be found in Rubinstein et al. (2008).

2.3 Exact variable selection

After removing many of the irrelevant variables have been removed using Algorithm 1, we are left with the variables in the set $\mathcal{M}^{(\hat{s})}$, whose size is smaller

than the sample size n . These variables are candidates for the relevant variables for each of the regressions. Now, we can address the problem of estimating the regression coefficients and recovering the exact support of \mathbf{B} using a lower dimensional selection procedure. In this paper, we use the adaptive Lasso as a lower dimensional selection procedure, which was shown to have oracle properties (Zou, 2006). The ALasso solves the penalized least square problem

$$\hat{\beta}_t = \underset{\beta_t \in \mathbb{R}^{\hat{s}}}{\operatorname{argmin}} \|\mathbf{y}_t - \mathbf{X}_{t, \mathcal{M}^{(\hat{s})}} \beta_t\|_2^2 + \lambda \sum_{j \in \mathcal{M}^{(\hat{s})}} w_j |\beta_{t,j}|, \quad (7)$$

where $(w_j)_{j \in \mathcal{M}^{(\hat{s})}}$ is a vector of known weight and λ is a tuning parameter. Usually, the weights are defined as $w_j = 1/|\hat{\beta}_{t,j}|$ where $\hat{\beta}_t$ is a \sqrt{n} -consistent estimator of β_t . In a low dimensional setting, we know from Huang et al. (2008) that the adaptive Lasso can recover the exactly the relevant variables. Therefore, we can use the ALasso on each output separately to recover the exact support of \mathbf{B} . However, in order to ensure that the exact support of \mathbf{B} is recovered with high probability, we need to have that the total number of tasks is $o(n)$. The exact support recovery of \mathbf{B} is established using the union bound over different tasks, therefore we need the number of tasks to remain relatively small in comparison to the sample size n . However, simulation results presented in § ref-sec:simulation show that the ALasso procedure succeeds in the exact support recovery even when the number of tasks are much larger than the sample size, which indicates that our theoretical considerations could be improved. Figure 1 illustrates the two step procedure.

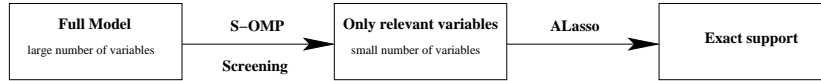


Figure 1: Framework for exact support recovery

Remark: We point out that solving the multi-task problem defined in (2) can be efficiently done on the reduced set of variables, but it is not obvious how to obtain the estimate of the exact support using (2). In Section 4.1, our numerical studies show that the ALasso applied to the reduced set of variables can be used to estimate the exact support of \mathbf{B} .

3 Theory

In this section we state conditions under which Algorithm 1 is screening consistent, i.e.,

$$\mathbb{P}[\exists k \in \{0, 1, \dots, n-1\} : \mathcal{M}_* \subseteq \mathcal{M}^{(k)}] \rightarrow 1, \text{ as } n \rightarrow \infty. \quad (8)$$

Furthermore, we also show that the model selected using the modified BIC criterion contains all the relevant variables, i.e.,

$$\mathbb{P}[\mathcal{M}_* \subseteq \mathcal{M}^{(\hat{s})}] \rightarrow 1, \text{ as } n \rightarrow \infty. \quad (9)$$

Note that we can choose trivially $\mathcal{M}^{(n)}$ since it holds that $\mathcal{M}_* \subseteq \mathcal{M}^{(n)}$. However, we will be able to prove that \hat{s} chosen by the modified BIC criterion is much smaller than the sample size n .

3.1 Assumptions

Before we state the theorem characterizing the performance of the S-OMP, we give some technical conditions that are needed for our analysis.

A1: The random noise vectors $\epsilon_1, \dots, \epsilon_T$ are independent Gaussian with zero mean and covariance matrix $\sigma^2 \mathbf{I}_{n \times n}$.

A2: Each row of the design matrix \mathbf{X}_t is IID Gaussian with zero mean and covariance matrix Σ_t . Furthermore, there exist two positive constants $0 < \phi_{\min} < \phi_{\max} < \infty$ such that

$$\phi_{\min} \leq \min_{t \in [T]} \Lambda_{\min}(\Sigma_t) \leq \max_{t \in [T]} \Lambda_{\max}(\Sigma_t) \leq \phi_{\max}. \quad (10)$$

A3: The true regression coefficients are bounded, i.e., there exists a positive constant C_β such that $\|\mathbf{B}\|_{2,1} \leq C_\beta$. Furthermore, the norm of any non-zero row of the matrix \mathbf{B} is bounded away from zero, that is, there exist positive constants c_β and δ_{\min} such that

$$T^{-1} \min_{j \in \mathcal{M}_*} \sum_{t \in [T]} \beta_{t,j}^2 \geq c_\beta n^{-\delta_{\min}}.$$

A4: There exist positive constants C_s, C_p, δ_s and δ_p such that $|\mathcal{M}_*| \leq C_s n^{\delta_s}$ and $\log(p) \leq C_p n^{\delta_p}$.

The normality condition **A1** is assumed here only to facilitate presentation of theoretical results, as is commonly assumed in literature, (e.g., Zhang and Huang, 2008; Fan and Lv, 2008). The normality assumption can be avoided at the cost of more technical proofs, e.g., Lounici et al. (2009), where the main technical difficulty is showing that the concentration properties still hold. Under the condition **A2** we will be able to show that the empirical covariance matrix satisfies the sparse eigenvalue condition (see Lemma 3) with probability tending to one. The assumption that the rows of the design are Gaussian can be easily relaxed to the case when the rows are sub-Gaussian, without any technical difficulties in proofs, since we would still obtain exponential bounds on the tail probabilities. The condition **A3** states that the regression coefficients are bounded, which is a technical condition likely to be satisfied in practice. Furthermore, it is assumed that the row norms of $\mathbf{B}_{\mathcal{M}_*}$ do not decay to zero too fast or, otherwise, they would not be distinguishable from noise. The condition is not too restrictive, e.g., if every non-zero coefficient is bounded away from zero by a constant, the condition **A3** is trivially satisfied with $\delta_{\min} = 0$. However, we allow the coefficients of the relevant variables to get smaller as the sample size increases and

still guarantee that the relevant variable will be identified. The condition **A4** sets the upper bound on the number of relevant variables and the total number of variables. While the total number of variables can diverge to infinity much faster than the sample size, the number of relevant variables needs to be smaller than the sample size. Conditions **A3** and **A4** implicitly relate different outputs and control the number of non-zero coefficients shared between different outputs. Intuitively, if the upper bound in **A4** on the size of \mathcal{M}_* is large, this immediately implies that the constant C_β in **A3** should be large as well, since otherwise there would exist a row of \mathbf{B} whose ℓ_2 norm would be too small to be detected by Algorithm 1.

3.2 Screening consistency

Our first results states that after a small number of iterations, compared to the dimensionality p , the S-OMP procedure will include all the relevant variables.

Theorem 1. *Assume the model in (3) and that the conditions **A1-A4** are satisfied. Furthermore, assume that*

$$\frac{n^{1-6\delta_s-6\delta_{\min}}}{\max\{\log(p), \log(T)\}} \rightarrow \infty, \text{ as } n \rightarrow \infty. \quad (11)$$

Then there exists a number $m_{\max}^ = m_{\max}^*(n)$, so that in m_{\max}^* steps of S-OMP iteration, all the relevant variables are included in the model, i.e., as $n \rightarrow \infty$*

$$\mathbb{P}[\mathcal{M}_* \subseteq \mathcal{M}^{(m_{\max}^*)}] \geq 1 - C_1 \exp\left(-C_2 \frac{n^{1-6\delta_s-6\delta_{\min}}}{\max\{\log(p), \log(T)\}}\right), \quad (12)$$

for some positive constants C_1 and C_2 . The exact value of m_{\max}^ is given as*

$$m_{\max}^* = \lfloor 2^4 \phi_{\min}^{-2} \phi_{\max} C_\beta^2 C_s^2 c_\beta^{-2} n^{2\delta_s+2\delta_{\min}} \rfloor. \quad (13)$$

Remarks: Under the assumptions of Theorem 1, $m_{\max}^* \leq n - 1$, so that the procedure effectively reduces the dimensionality below the sample size. From the proof of the theorem, it is clear how multiple outputs help to identify the relevant variables. The crucial quantity in identifying all relevant variables is the minimum non-zero row norm of \mathbf{B} , which allows us to identify weak variables if they are relevant for a large number of outputs even though individual coefficients may be small. It should be noted that the main improvement over the ordinary forward regression is in the seize of the signal that can be detected, as defined in **A3** and **A4**.

Theorem 1 guarantees that one of the sets $\{\mathcal{M}^{(k)}\}$ will contain all relevant variables, with high probability. However, it is of practical importance to select of one set in the collection that contains all relevant variables and does not have too many irrelevant ones. Our following theorem shows that the modified BIC criterion can be used for this purpose, that is, the set $\mathcal{M}^{(\hat{s})}$ is screening consistent.

Theorem 2. Assume that the conditions of Theorem 1 are satisfied. Let

$$\hat{s} = \underset{k \in \{0, \dots, n-1\}}{\operatorname{argmin}} \operatorname{BIC}(\mathcal{M}^{(k)}) \quad (14)$$

be the index of the model selected by optimizing the modified BIC criterion. Then, as $n \rightarrow \infty$

$$\mathbb{P}[\mathcal{M}_* \subseteq \mathcal{M}^{(\hat{s})}] \rightarrow 1. \quad (15)$$

Combining results from Theorem 1 and Theorem 2, we have shown that the S-OMP procedure is screening consistent and can be applied to problems where the dimensionality of the problem p is exponential in the number of observed samples. In the next section, we also show that the S-OMP has great empirical performance.

4 Numerical studies

In this section we perform simulation studies on an extensive number synthetic data sets. Furthermore, we demonstrate the application of the procedure on the genome-wide association mapping problem.

4.1 Simulation studies

We conduct an extensive number of numerical studies to evaluate the finite sample performance of the S-OMP. We consider three procedures that perform estimation on individuals outputs: Sure Independence Screening (SIS), Iterative SIS (ISIS) (Fan and Lv, 2008), and the OMP, for comparison purposes. The evaluation is done on the model in (1). SIS and ISIS are used to select a subset of variables and then the ALasso is used to further refine the selection. We denote this combination as SIS-ALasso and ISIS-ALasso. The size of the model selected by SIS is fixed as $n-1$, while the ISIS selects $\lfloor n/\log(n) \rfloor$ variables in each of the $\lfloor \log(n) - 1 \rfloor$ iterations. From the screened variables, the final model is selected using the ALasso, together with the BIC criterion (6) to determine the penalty parameter λ . The number of variables selected by the OMP is determined using the BIC criterion, however, we do not further refine the selected variables using the ALasso, since from the numerical studies in Wang (2009) it was observed that the further refinement does not result in improvement. The S-OMP is used to reduce the dimensionality below the sample size jointly using the regression outputs. Next, the ALasso is used on each of the outputs to further perform the estimation. This combination is denoted SOMP-ALasso.

Let $\hat{\mathbf{B}} = [\hat{\boldsymbol{\beta}}_1, \dots, \hat{\boldsymbol{\beta}}_T] \in \mathbb{R}^{p \times T}$ be an estimate obtained by one of the estimation procedures. We evaluate the performance averaged over 200 simulation runs. Let $\hat{\mathbb{E}}_n$ denote the empirical average over the simulation runs. We measure the size of the union support $\hat{S} = S(\hat{\mathbf{B}}) := \{j \in [p] : \|\hat{\mathbf{B}}_j\|_2^2 > 0\}$. Next, we estimate the probability that the screening property is satisfied $\hat{\mathbb{E}}_n[\mathbb{I}\{\mathcal{M}_* \subseteq S(\hat{\mathbf{B}})\}]$, which we call coverage probability. For the union support, we define fraction of correct zeros $(p-s)^{-1}\hat{\mathbb{E}}_n[|S(\hat{\mathbf{B}})^C \cap \mathcal{M}_*^C|]$, fraction of incorrect zeros

$s^{-1}\hat{\mathbb{E}}_n[|S(\hat{\mathbf{B}})^C \cap \mathcal{M}_*|]$ and fraction of correctly fitted $\hat{\mathbb{E}}_n[\mathbb{I}\{\mathcal{M}_* = S(\hat{\mathbf{B}})\}]$ to measure the performance of different procedures. Similar quantities are defined for the exact support recovery. In addition, we measure the estimation error $\hat{\mathbb{E}}_n[\|\mathbf{B} - \hat{\mathbf{B}}\|_2^2]$ and the prediction performance on the test set. On the test data $\{\mathbf{x}_i^*, \mathbf{y}_i^*\}_{i \in [n]}$, we compute

$$R^2 = 1 - \frac{\sum_{i \in [n]} \sum_{t \in [T]} (y_{t,i}^* - (\mathbf{x}_{t,i}^*)' \hat{\boldsymbol{\beta}}_t)^2}{\sum_{i \in [n]} \sum_{t \in [T]} (y_{t,i}^* - \bar{y}_t^*)^2}, \quad (16)$$

where $\bar{y}_t^* = n^{-1} \sum_{i \in [n]} y_{t,i}^*$.

The following simulation studies are used to comparatively assess the numerical performance of the procedures. Due to space constraints, tables with detailed numerical results are given in the appendix. In this section, we outline main findings.

Simulation 1: [Model with uncorrelated variables] The following toy model is based on the simulation I in Fan and Lv (2008) with $(n, p, s, T) = (400, 20000, 18, 500)$. Each \mathbf{x}_i is drawn independently from a standard multivariate normal distribution, so that the variables are mutually independent. For $j \in [s]$ and $t \in [T]$, the non-zero coefficients of \mathbf{B} are given as $\beta_{t,j} = (-1)^u (4n^{-1/2} \log n + |z|)$, where $u \sim \text{Bernoulli}(0.4)$ and $z \sim \mathcal{N}(0, 1)$. The number of non-zero elements in \mathbf{B}_j is given as a parameter $T_{\text{non-zero}} \in \{500, 300, 100\}$. The positions of non-zero elements are chosen uniformly at random from $[T]$. The noise is Gaussian with the standard deviation σ set to control the signal-to-noise ratio (SNR). SNR is defined as $\text{Var}(\mathbf{x}\boldsymbol{\beta}) / \text{Var}(\boldsymbol{\epsilon})$ and we vary $\text{SNR} \in \{15, 10, 5, 1\}$.

Simulation 2: [Changing the number of non-zero elements in \mathbf{B}_j] The following scenario is used to evaluate the performance of the methods as the number of non-zero elements in a row of \mathbf{B} varies. We set $(n, p, s) = (100, 500, 10)$ and vary the number of outputs $T \in \{500, 750, 1000\}$. For each number of outputs T , we vary $T_{\text{non-zero}} \in \{0.8T, 0.5T, 0.2T\}$. The samples \mathbf{x}_i and regression coefficients \mathbf{B} are given as in Simulation 1, i.e., \mathbf{x}_i is drawn from a multivariate standard normal distribution and the non-zero coefficients \mathbf{B} are given as $\beta_{t,j} = (-1)^u (4n^{-1/2} \log n + |z|)$, where $u \sim \text{Bernoulli}(0.4)$ and $z \sim \mathcal{N}(0, 1)$. The noise is Gaussian, with the standard deviation defined through the SNR, which varies in $\{10, 5, 1\}$.

Simulation 3: [Model with the decaying correlation between variables] The following model is borrowed from Wang (2009). We assume a correlation structure between variables given as $\text{Var}(\mathbf{X}_{j_1}, \mathbf{X}_{j_2}) = \rho^{|j_1 - j_2|}$, where $\rho \in \{0.2, 0.5, 0.7\}$. This correlation structure appears naturally among ordered variables. We set $(n, p, s, T) = (100, 5000, 3, 150)$ and $T_{\text{non-zero}} = 80$. The relevant variables are at positions $(1, 4, 7)$ and non-zero coefficients are given as 3, 1.5 and 2 respectively. The SNR varies in $\{10, 5, 1\}$. A heat map of the correlation matrix between different covariates is given in Figure 2.

Simulation 4: [Model with the block-compound correlation structure] The following model assumes a block compound correlation structure. For a parameter ρ , the correlation between two variables \mathbf{X}_{j_1} and \mathbf{X}_{j_2} is given as ρ , ρ^2 or ρ^3 when $|j_1 - j_2| \leq 10$, $|j_1 - j_2| \in (10, 20]$ or $|j_1 - j_2| \in (20, 30]$ and it is set

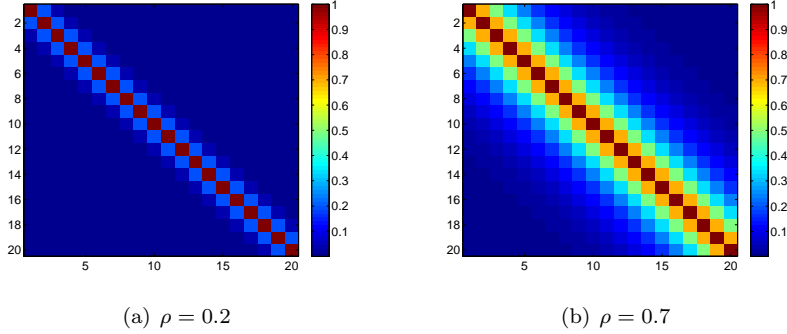


Figure 2: Visualization of the correlation matrix in Simulation 3. Only an upper left corner is presented corresponding to 20 of the 5000 variables.

to 0 otherwise. We set $(n, p, s, T) = (150, 4000, 8, 150)$, $T_{\text{non-zero}} = 80$ and the parameter $\rho \in \{0.2, 0.5\}$. The relevant variables are located at positions 1, 11, 21, 31, 41, 51, 61, 71 and 81, so that each block of highly correlated variables has exactly one relevant variable. The values of relevant coefficients are given as in Simulation 1. The noise is Gaussian and the SNR varies in $\{10, 5, 1\}$. A heatmap of the correlation matrix between different covariates is given in Figure 3.

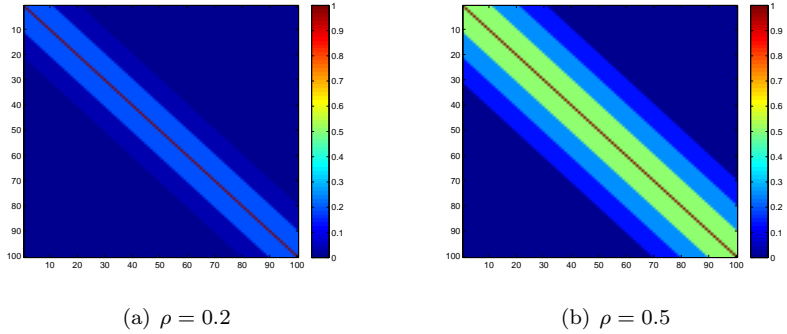


Figure 3: Visualization of the correlation matrix in Simulation 4. Only an upper left corner is presented corresponding to 100 of the 4000 variables.

Simulation 5: [Model with a 'masked' relevant variable] This model represents a difficult setting. It is modified from Wang (2009). We set $(n, p, s, T) = (200, 10000, 5, 500)$. The number of non-zero elements in each row varies is $T_{\text{non-zero}} \in \{400, 250, 100\}$. For $j \in [s]$ and $t \in [T]$, the non-zero elements equal $\beta_{t,j} = 2j$. Each row of \mathbf{X} is generated as follows. Draw independently \mathbf{z}_i and \mathbf{z}'_i from a p -dimensional standard multivariate normal distribution. Now, $x_{ij} = (z_{ij} + z'_{ij})/\sqrt{2}$ for $j \in [s]$ and $x_{ij} = (z_{ij} + \sum_{j' \in [s]} z_{ij'})/2$ for $j \in [p] \setminus [s]$. Now, $\text{Corr}(x_{i,1}, y_{t,i})$ is much smaller than $\text{Corr}(x_{i,j}, y_{t,i})$ for $j \in [p] \setminus [s]$, so that it

becomes difficult to select variable 1. The variable 1 is 'masked' with the noisy variables. This setting is difficult for screening procedures as they take into consideration only marginal information. The noise is Gaussian with standard deviation $\sigma \in \{1.5, 2.5, 4.5\}$.

In the next section we summarize results of our experimental findings. Our simulation setting transitions from a simple scenario considered in Simulation 1 towards a challenging one in Simulation 5. Simulation 1 is adopted from Fan and Lv (2008) as a toy model on which all algorithms should work well. Simulation 2 examines the influence of the number of non-zero elements in a relevant row of the matrix \mathbf{B} . We expect that Algorithm 1 will outperform procedures that perform estimation on individual outputs when $T_{\text{non-zero}}$ is large, while when $T_{\text{non-zero}}$ is small the single-task screening procedures should have an advantage. Our intuition is also supported by recent results of Kolar et al. (2010). Simulations 3 and 4 represent more challenging situations with structured correlation that naturally appears in many data sets, for example, a correlation between gene measurements that are closely located on a chromosome. Finally Simulation 5 is constructed in such a way that procedures which use only marginal information are going to include irrelevant variables before relevant ones.

4.2 Results of simulations

Tables giving detailed results of the above described simulations are given in the Appendix. In this section, we outline main findings and reproduce some parts of the tables that we think are insightful.

Table 1 shows parts of the results for simulation 1. We can see that all methods perform well in the setting when the input variables are mutually uncorrelated and the SNR is high. Note that even though the variables are uncorrelated, the sample correlation between variables can be quite high due to large p and small n , which can result in selection of spurious variables. As we can see from the table, comparing to SIS, ISIS and OMP, the S-OMP is able to select the correct union support, while the procedures that select variables based on different outputs separately also include additional spurious variables into the selection. Furthermore, we can see that the S-OMP-ALasso procedure does much better on the problem of exact support recovery compared to the other procedures. The first simulations suggests that somewhat higher computational cost of the S-OMP procedure can be justified by the improved performance on the problem of union and exact support recovery as well as on the error in the estimated coefficients.

Table 2 shows parts of the results for simulation 2. In this simulation we measured the performance of estimation procedures as the amount of shared input variables between different outputs varies. The parameter $T_{\text{non-zero}}$ controls the amount of information that is shared between different tasks as defined in the previous subsection. In particular, the parameter controls the number of non-zero elements in a row of the matrix \mathbf{B} corresponding to a relevant variable. When the number of non-zero elements is high, a variable is relevant to many

Table 1: Results for simulation 1 with parameters $(n, p, s, T) = (500, 20000, 18, 500)$, $T_{\text{non-zero}} = 500$

	Method name	Prob. (%) of $\mathcal{M}_* \subseteq \hat{S}$	Fraction (%) of Correct zeros	Fraction (%) of Incorrect zeros	Fraction (%) of $\mathcal{M}_* = \hat{S}$	$ \hat{S} $	Est. error $\ \mathbf{B} - \hat{\mathbf{B}}\ _2^2$	Test error R^2
SNR = 15								
Union Support	SIS-ALASSO	100.0	100.0	0.0	10.0	20.2	-	-
	ISIS-ALASSO	100.0	100.0	0.0	18.0	19.6	-	-
	OMP	100.0	100.0	0.0	0.0	23.9	-	-
	S-OMP	100.0	100.0	0.0	100.0	18.0	-	-
	S-OMP-ALASSO	100.0	100.0	0.0	100.0	18.0	-	-
Exact Support	SIS-ALASSO	0.0	100.0	0.7	0.0	8940.5	0.97	0.93
	ISIS-ALASSO	100.0	100.0	0.0	18.0	9001.6	0.33	0.93
	OMP	100.0	100.0	0.0	0.0	9005.9	0.20	0.93
	S-OMP-ALASSO	100.0	100.0	0.0	100.0	9000.0	0.20	0.93

Table 2: Results for simulation 2 with parameters $(n, p, s, T) = (200, 5000, 10, 1000)$, $T_{\text{non-zero}} = 200$

	Method name	Prob. (%) of $\mathcal{M}_* \subseteq \hat{S}$	Fraction (%) of Correct zeros	Fraction (%) of Incorrect zeros	Fraction (%) of $\mathcal{M}_* = \hat{S}$	$ \hat{S} $	Est. error $\ \mathbf{B} - \hat{\mathbf{B}}\ _2^2$	Test error R^2
SNR = 5								
Union Support	SIS-ALASSO	100.0	100.0	0.0	100.0	10.0	-	-
	ISIS-ALASSO	100.0	100.0	0.0	100.0	10.0	-	-
	OMP	100.0	97.4	0.0	0.0	139.6	-	-
	S-OMP	100.0	100.0	0.0	100.0	10.0	-	-
	S-OMP-ALASSO	100.0	100.0	0.0	100.0	10.0	-	-
Exact Support	SIS-ALASSO	100.0	100.0	0.0	100.0	2000.0	0.04	0.72
	ISIS-ALASSO	100.0	100.0	0.0	100.0	2000.0	0.04	0.72
	OMP	100.0	100.0	0.0	0.0	2131.6	0.05	0.71
	S-OMP-ALASSO	100.0	100.0	0.0	100.0	2000.0	0.03	0.72

Table 3: Results for simulation 3 with parameters $(n, p, s, T) = (100, 5000, 3, 150)$, $T_{\text{non-zero}} = 80$, $\rho = 0.5$

	Method name	Prob. (%) of $\mathcal{M}_* \subseteq \hat{S}$	Fraction (%) of Correct zeros	Fraction (%) of Incorrect zeros	Fraction (%) of $\mathcal{M}_* = \hat{S}$	$ \hat{S} $	Est. error $\ \mathbf{B} - \hat{\mathbf{B}}\ _2^2$	Test error R^2
SNR = 5								
Union Support	SIS-ALASSO	100.0	100.0	0.0	97.0	3.0	-	-
	ISIS-ALASSO	100.0	100.0	0.0	96.0	3.0	-	-
	OMP	100.0	99.8	0.0	0.0	19.6	-	-
	S-OMP	100.0	100.0	0.0	100.0	3.0	-	-
	S-OMP-ALASSO	100.0	100.0	0.0	100.0	3.0	-	-
Exact Support	SIS-ALASSO	60.0	100.0	0.2	57.0	239.5	0.10	0.61
	ISIS-ALASSO	84.0	100.0	0.1	80.0	239.8	0.08	0.61
	OMP	100.0	100.0	0.0	0.0	256.6	0.06	0.61
	S-OMP-ALASSO	100.0	100.0	0.0	100.0	240.0	0.03	0.62

Table 4: Results of simulation 4 with parameters $(n, p, s, T) = (150, 4000, 8, 150)$, $T_{\text{non-zero}} = 80$, $\rho = 0.5$

		Prob. (%) of	Fraction (%) of	Fraction (%) of	Fraction (%) of		Est. error	Test error
Method name		$\mathcal{M}_* \subseteq \hat{S}$	Correct zeros	Incorrect zeros	$\mathcal{M}_* = \hat{S}$	$ \hat{S} $	$\ \mathbf{B} - \hat{\mathbf{B}}\ _2^2$	R^2
SNR = 10								
Union Support	SIS-ALASSO	100.0	100.0	0.0	100.0	8.0	-	-
	ISIS-ALASSO	100.0	100.0	0.0	97.0	8.0	-	-
	OMP	100.0	99.9	0.0	2.0	11.7	-	-
	S-OMP	100.0	100.0	0.0	100.0	8.0	-	-
	S-OMP-ALASSO	100.0	100.0	0.0	100.0	8.0	-	-
Exact Support	SIS-ALASSO	35.0	100.0	1.4	35.0	631.3	0.55	0.88
	ISIS-ALASSO	100.0	100.0	0.0	97.0	640.0	0.14	0.89
	OMP	100.0	100.0	0.0	2.0	643.7	0.10	0.89
	S-OMP-ALASSO	100.0	100.0	0.0	100.0	640.0	0.09	0.89

Table 5: Results of simulation 5 with parameters $(n, p, s, T) = (200, 10000, 5, 500)$, $T_{\text{non-zero}} = 400$

		Prob. (%) of	Fraction (%) of	Fraction (%) of	Fraction (%) of		Est. error	Test error
Method name		$\mathcal{M}_* \subseteq \hat{S}$	Correct zeros	Incorrect zeros	$\mathcal{M}_* = \hat{S}$	$ \hat{S} $	$\ \mathbf{B} - \hat{\mathbf{B}}\ _2^2$	R^2
$\sigma = 1.5$								
Union Support	SIS-ALASSO	53.0	99.6	9.4	0.0	41.1	-	-
	ISIS-ALASSO	100.0	99.8	0.0	0.0	28.1	-	-
	OMP	100.0	99.9	0.0	12.0	10.0	-	-
	S-OMP	100.0	100.0	0.0	44.0	5.6	-	-
	S-OMP-ALASSO	100.0	100.0	0.0	100.0	5.0	-	-
Exact Support	SIS-ALASSO	0.0	100.0	68.9	0.0	936.0	84.66	0.66
	ISIS-ALASSO	0.0	100.0	16.2	0.0	1791.9	5.80	0.96
	OMP	100.0	100.0	0.0	12.0	2090.3	0.06	0.99
	S-OMP-ALASSO	100.0	100.0	0.0	100.0	2000.0	0.05	0.99

tasks and we say that outputs overlap. In this setting the S-OMP procedure is expected to outperform the other methods, however, when $T_{\text{non-zero}}$ is low, the noise coming from the tasks for which the variable is irrelevant can actually harm the performance. The table shows results when the overlap of shared variables is small, that is, a relevant variable is only relevant for 10% of outputs. As one could expect, the S-OMP procedure does as well as other procedures. This is not surprising, since the amount of shared information between different outputs is limited. Therefore, if one expects little variable sharing across different outputs, using the SIS or ISIS may results in similar accuracy, but an improved computational efficiency. It is worth pointing out that in our simulations, the different tasks are correlated since the same design \mathbf{X} used for all tasks. However, we expect the same qualitative results even under the model given in equation (3) where different tasks can have different designs \mathbf{X}_t and the outputs are uncorrelated.

Simulation 3 represents a situation that commonly occur in nature, where there is an ordering among input variables and the correlation between variables decays as the distance between variables increase. The model in simulation 4 is a modification of the model in simulation 3 where the variables are grouped and there is some correlation between different groups. Table 3 gives results for simulation 3 for the parameter $\rho = 0.5$. In this setting, the S-OMP performs much better than the other procedures. The improvement becomes more pronounced with increase of the correlation parameter ρ . Similar behavior is observed in simulation 4 as well, see table 4. Results of simulation 5, given in Table 5, further reinforce our intuition that the S-OMP procedure does well even on problems with high-correlation between the set of relevant input variables and the set of irrelevant ones.

To further compare the performance of the S-OMP procedure to the SIS, we explore the minimum number of iterations needed for the algorithm to include all the relevant variables into the selected model. From our limited numerical experience, we note that the simulation parameters do not affect the number of iterations for the S-OMP procedure. This is unlike the SIS procedure, which occasionally requires a large number of steps before all the true variables are included, see Figure 3 in Fan and Lv (2008). We note that while the S-OMP procedure does include, in many cases, all the relevant variables before the irrelevant ones, the BIC criterion is not able to correctly select the number of variables to include, when the SNR is small. As a result, we can see the drop in performance as the SNR decreases.

4.3 Real data analysis

We demonstrate an application of the S-OMP to a genome-wide association mapping problem. The data were collected by our collaborator Judie Howrylak, M.D. at Harvard Medical School from 200 individuals that are suffering from asthma. For each individual, we have a collection of about ~ 350000 ge-

netic markers¹, which are called single nucleotide polymorphisms (SNPs), and a collection of 1424 gene expression measurements. The goal of this study is to identify a small number of SNPs that can help explain variations in gene expressions. Typically, this type of analysis is done by regressing each gene individually on the measured SNPs, however, since the data are very noisy, such an approach results selecting many variables. Our approach to this problem here is to regress a group of genes onto the SNPs instead. There has been some previous work on this problem Kim and Xing (2009), that considered regressing groups of genes onto SNPs, however, those approaches use variants of the estimation procedure given in Eq. (2), which is not easily scalable to the data we analyze here.

We use the spectral relaxation of the k-means clustering Zha et al. (2001) to group 1424 genes into 48 clusters according to their expression values, so that the minimum, maximum and median number of genes per cluster is 4, 90 and 19, respectively. The number of clusters was chosen somewhat arbitrarily, based on the domain knowledge of the medical experts. The main idea behind the clustering is that we want to identify genes that belong to the same regulatory pathway since they are more likely to be affected with the same SNPs. Instead of clustering, one may use prior knowledge to identify interesting groups of genes. Next, we want to use the S-OMP procedure to identify relevant SNPs for each of the gene clusters. Since, we do not have the ground truth for the data set, we use predictive power on the test set and the size of estimated models to access their quality. We randomly split the data into a training set of size 170 and a testing set of size 30 and report results over 500 runs. We compute the R^2 coefficient on the test set defined as $1 - 30^{-1}T^{-1} \sum_{t \in [T]} \|\mathbf{y}_{t,\text{test}} - \mathbf{X}_{t,\text{test}}\hat{\beta}_t\|_2^2$ (because the data has been normalized).

Due to space constraints, we give results on few clusters in Table 6 and note that, qualitatively, the results do not vary much between different clusters. While the fitted models have limited predictive performance, which results from highly noisy data, we observe that the S-OMP is able to identify on average one SNP per gene cluster that is related to a large number of genes. Other methods, while having a similar predictive performance, select a larger number of SNPs which can be seen from the size of the union support. On this particular data set, the S-OMP seems produce results that are more interpretable from a specialist’s points of view. Further investigation needs to be done to verify the biological significance of the selected SNPs, however, the details of such an analysis are going to be reported elsewhere.

5 Conclusions

In this work, we analyze the Simultaneous Orthogonal Matching Pursuit as a method for variable selection in an ultra-high dimensional space. We prove that the S-OMP is screening consistent and provide a practical way to select the

¹These markers were preprocessed, by imputing missing values and removing duplicate SNPs that were perfectly correlated with other SNPs.

Table 6: Results on the asthma data

	Method name	Union support	R^2
Cluster 9 Size = 18	SIS-ALASSO	18.0 (1.0)	0.178 (0.006)
	OMP	17.5 (2.9)	0.167 (0.002)
	S-OMP	1.0 (0.0)	0.214 (0.005)
Cluster 16 Size = 31	SIS-ALASSO	31.0 (1.0)	0.160 (0.007)
	OMP	29.0 (1.8)	0.165 (0.002)
	S-OMP	1.0 (0.0)	0.209 (0.005)
Cluster 17 Size = 19	SIS-ALASSO	18.5 (0.9)	0.173 (0.006)
	OMP	19.5 (0.8)	0.146 (0.003)
	S-OMP	1.0 (0.0)	0.184 (0.004)
Cluster 19 Size = 17	SIS-ALASSO	17.0 (1.2)	0.270 (0.017)
	OMP	11.0 (4.1)	0.213 (0.008)
	S-OMP	1.0 (0.0)	0.280 (0.017)
Cluster 22 Size = 34	SIS-ALASSO	34.0 (0.9)	0.153 (0.005)
	OMP	30.0 (7.3)	0.142 (0.000)
	S-OMP	1.0 (0.0)	0.145 (0.002)
Cluster 23 Size = 35	SIS-ALASSO	35.0 (0.9)	0.238 (0.018)
	OMP	33.0 (9.9)	0.208 (0.009)
	S-OMP	1.0 (0.0)	0.229 (0.014)
Cluster 24 Size = 28	SIS-ALASSO	28.0 (1.0)	0.123 (0.003)
	OMP	28.0 (2.6)	0.114 (0.001)
	S-OMP	1.0 (0.0)	0.129 (0.003)
Cluster 32 Size = 15	SIS-ALASSO	15.0 (0.9)	0.188 (0.010)
	OMP	10.0 (2.6)	0.211 (0.006)
	S-OMP	1.0 (0.0)	0.215 (0.008)
Cluster 36 Size = 33	SIS-ALASSO	34.0 (1.4)	0.147 (0.005)
	OMP	29.0 (5.3)	0.157 (0.002)
	S-OMP	1.0 (0.0)	0.168 (0.004)
Cluster 37 Size = 19	SIS-ALASSO	19.0 (0.9)	0.207 (0.015)
	OMP	22.0 (2.5)	0.175 (0.006)
	S-OMP	1.0 (0.0)	0.235 (0.014)
Cluster 39 Size = 24	SIS-ALASSO	24.0 (0.9)	0.131 (0.006)
	OMP	27.0 (1.9)	0.141 (0.003)
	S-OMP	1.0 (0.0)	0.160 (0.005)
Cluster 44 Size = 35	SIS-ALASSO	35.0 (0.9)	0.177 (0.010)
	OMP	26.5 (6.6)	0.183 (0.005)
	S-OMP	1.0 (0.0)	0.170 (0.011)
Cluster 49 Size = 23	SIS-ALASSO	23.0 (1.0)	0.124 (0.004)
	OMP	23.0 (1.2)	0.140 (0.000)
	S-OMP	1.0 (0.0)	0.159 (0.004)

number of steps in the procedure using the modified Bayesian information criterion. Our limited numerical experience shows that the method performs well in practice and that the joint estimation from multiple outputs often outperforms methods that use one regression output at the time. Furthermore, we can see the S-OMP procedure as way to improve the variable selection properties of the SIS without having to solve a costly complex optimization procedure in Eq. (2), therefore, balancing the computational costs and the estimation accuracy.

6 Appendix

6.1 Proof of Theorem 1

Under the assumptions of the theorem, the number of relevant variables s is relatively small compared to the sample size n . The proof strategy can be outlined as follows: i) we are going to show that, with high probability, at least one relevant variable is going to be identified within the following m_{one}^* steps, conditioning on the already selected variables $\mathcal{M}^{(k)}$ and this holds uniformly for all k ; ii) we can conclude that all the relevant variables are going to be selected within $m_{\text{max}}^* = sm_{\text{one}}^*$ steps. Exact values for m_{one}^* and m_{max}^* are given below. Without loss of generality, we analyze the first step of the algorithm, i.e., we show that the first relevant variable is going to be selected within the first m_{one}^* steps.

Assume that in the first $m_{\text{one}}^* - 1$ steps, there were no relevant variables selected. Assuming that the variable selected in the m_{one}^* -th step is still an irrelevant one, we will arrive at a contradiction, which shows that at least one relevant variable has been selected in the first m_{one}^* steps. For any step k , the reduction of the squared error is given as

$$\Delta(k) := \text{RSS}(k-1) - \text{RSS}(k) = \sum_t \|\mathbf{H}_{t, \hat{f}_k}^{(k)} (\mathbf{I}_{n \times n} - \mathbf{H}_{t, \mathcal{M}^{(k)}}) \mathbf{y}_t\|_2^2 \quad (17)$$

with $\mathbf{H}_{t, j}^{(k)} = \mathbf{X}_{t, j}^{(k)} \mathbf{X}_{t, j}^{(k)'} \|\mathbf{X}_{t, j}^{(k)}\|^{-2}$ and $\mathbf{X}_{t, j}^{(k)} = (\mathbf{I}_{n \times n} - \mathbf{H}_{t, \mathcal{M}^{(k)}}) \mathbf{X}_{t, j}$. We are interested in the quantity $\sum_{k=1}^{m_{\text{one}}^*} \Delta(k)$, when all the selected variables \hat{f}_k (see Algorithm 1) belong to $[p] \setminus \mathcal{M}_*$.

In what follows, we will derive a lower bound for $\Delta(k)$. We perform our analysis on the event

$$\mathcal{E} = \left\{ \min_{t \in [T]} \min_{\mathcal{M} \subseteq [p], |\mathcal{M}| \leq m_{\text{max}}^*} \Lambda_{\min}(\hat{\Sigma}_{\mathcal{M}}) \geq \phi_{\min}/2 \right\} \cap \left\{ \max_{t \in [T]} \max_{\mathcal{M} \subseteq [p], |\mathcal{M}| \leq m_{\text{max}}^*} \Lambda_{\max}(\hat{\Sigma}_{\mathcal{M}}) \leq 2\phi_{\max} \right\}. \quad (18)$$

From the definition of \hat{f}_k , we have

$$\begin{aligned}
\Delta(k) &\geq \max_{j \in \mathcal{M}_*} \sum_t \|\mathbf{H}_{t,j}^{(k)} (\mathbf{I}_{n \times n} - \mathbf{H}_{t,\mathcal{M}^{(k)}}) \mathbf{y}_t\|_2^2 \\
&\geq \max_{j \in \mathcal{M}_*} \left(\sum_t \|\mathbf{H}_{t,j}^{(k)} (\mathbf{I}_{n \times n} - \mathbf{H}_{t,\mathcal{M}^{(k)}}) \mathbf{X}_{t,\mathcal{M}_*} \boldsymbol{\beta}_{t,\mathcal{M}_*}\|_2^2 \right. \\
&\quad \left. - \sum_t \|\mathbf{H}_{t,j}^{(k)} (\mathbf{I}_{n \times n} - \mathbf{H}_{t,\mathcal{M}^{(k)}}) \boldsymbol{\epsilon}_t\|_2^2 \right) \\
&\geq \max_{j \in \mathcal{M}_*} \sum_t \|\mathbf{H}_{t,j}^{(k)} (\mathbf{I}_{n \times n} - \mathbf{H}_{t,\mathcal{M}^{(k)}}) \mathbf{X}_{t,\mathcal{M}_*} \boldsymbol{\beta}_{t,\mathcal{M}_*}\|_2^2 \\
&\quad - \max_{j \in \mathcal{M}_*} \sum_t \|\mathbf{H}_{t,j}^{(k)} (\mathbf{I}_{n \times n} - \mathbf{H}_{t,\mathcal{M}^{(k)}}) \boldsymbol{\epsilon}_t\|_2^2 \\
&= (I) - (II).
\end{aligned} \tag{19}$$

We deal with these two terms separately. Let $\mathbf{H}_{t,\mathcal{M}}^\perp = \mathbf{I}_{n \times n} - \mathbf{H}_{t,\mathcal{M}}$ denote the projection matrix. We have that the first term (I) is lower bounded by

$$\begin{aligned}
&\max_{j \in \mathcal{M}_*} \sum_t \|\mathbf{H}_{t,j}^{(k)} \mathbf{H}_{t,\mathcal{M}^{(k)}}^\perp \mathbf{X}_{t,\mathcal{M}_*} \boldsymbol{\beta}_{t,\mathcal{M}_*}\|_2^2 \\
&= \max_{j \in \mathcal{M}_*} \sum_t \|\mathbf{X}_{t,j}^{(k)}\|_2^{-2} |\mathbf{X}_{t,j}^{(k)'} \mathbf{H}_{t,\mathcal{M}^{(k)}}^\perp \mathbf{X}_{t,\mathcal{M}_*} \boldsymbol{\beta}_{t,\mathcal{M}_*}|^2 \\
&\geq \min_{t \in [T], j \in \mathcal{M}_*} \{\|\mathbf{X}_{t,j}^{(k)}\|_2^{-2}\} \max_{j \in \mathcal{M}_*} \sum_t |\mathbf{X}_{t,j}^{(k)'} \mathbf{H}_{t,\mathcal{M}^{(k)}}^\perp \mathbf{X}_{t,\mathcal{M}_*} \boldsymbol{\beta}_{t,\mathcal{M}_*}|^2 \\
&\geq \left\{ \max_{t \in [T], j \in \mathcal{M}_*} \|\mathbf{X}_{t,j}\|_2^2 \right\}^{-1} \max_{j \in \mathcal{M}_*} \sum_t |\mathbf{X}_{t,j}' \mathbf{H}_{t,\mathcal{M}^{(k)}}^\perp \mathbf{X}_{t,\mathcal{M}_*} \boldsymbol{\beta}_{t,\mathcal{M}_*}|^2,
\end{aligned} \tag{20}$$

where the last inequality follows from the fact that $\|\mathbf{X}_{t,j}\|_2 \geq \|\mathbf{X}_{t,j}^{(k)}\|_2$ and $\mathbf{X}_{t,j}^{(k)'} \mathbf{H}_{t,\mathcal{M}^{(k)}}^\perp = \mathbf{X}_{t,j}' \mathbf{H}_{t,\mathcal{M}^{(k)}}^\perp$. A simple calculation shows that

$$\begin{aligned}
&\sum_t \|\mathbf{H}_{t,\mathcal{M}^{(k)}}^\perp \mathbf{X}_{t,\mathcal{M}_*} \boldsymbol{\beta}_{t,\mathcal{M}_*}\|_2^2 \\
&= \sum_t \sum_{j \in \mathcal{M}_*} \boldsymbol{\beta}_{t,j} \mathbf{X}_{t,j} \mathbf{H}_{t,\mathcal{M}^{(k)}}^\perp \mathbf{X}_{t,\mathcal{M}_*} \boldsymbol{\beta}_{t,\mathcal{M}_*} \\
&\leq \sum_{j \in \mathcal{M}_*} \sqrt{\sum_t \beta_{t,j}^2} \sqrt{\sum_t (\mathbf{X}_{t,j} \mathbf{H}_{t,\mathcal{M}^{(k)}}^\perp \mathbf{X}_{t,\mathcal{M}_*} \boldsymbol{\beta}_{t,\mathcal{M}_*})^2} \\
&\leq \|\boldsymbol{\beta}\|_{2,1} \max_{j \in \mathcal{M}_*} \sqrt{\sum_t (\mathbf{X}_{t,j} \mathbf{H}_{t,\mathcal{M}^{(k)}}^\perp \mathbf{X}_{t,\mathcal{M}_*} \boldsymbol{\beta}_{t,\mathcal{M}_*})^2}.
\end{aligned} \tag{21}$$

Plugging (21) back into (20), the following lower bound is achieved

$$(I) \geq \left\{ \max_{t \in [T], j \in \mathcal{M}_*} \|\mathbf{X}_{t,j}\|_2^2 \right\}^{-1} \frac{(\sum_t \|\mathbf{H}_{t,\mathcal{M}^{(k)}}^\perp \mathbf{X}_{t,\mathcal{M}_*} \boldsymbol{\beta}_{t,\mathcal{M}_*}\|_2^2)^2}{\|\mathbf{B}\|_{2,1}^2}. \tag{22}$$

On the event \mathcal{E} , $\max_{t \in [T], j \in \mathcal{M}_*} \|\mathbf{X}_{t,j}\|_2^2 \leq 2n\phi_{\max}$. Since we have assumed that no additional relevant predictors have been selected by the procedure, it holds that $\mathcal{M}_* \not\subseteq \mathcal{M}^{(k)}$. This leads to

$$\sum_t \|\mathbf{H}_{t,\mathcal{M}^{(k)}}^\perp \mathbf{X}_{t,\mathcal{M}_*} \beta_{t,\mathcal{M}_*}\|_2^2 \geq 2^{-1} n \phi_{\min} \min_{j \in \mathcal{M}^*} \sum_{t \in [T]} \beta_{t,j}^2, \quad (23)$$

on the event \mathcal{E} . Using the Cauchy-Schwarz inequality, $\|\mathbf{B}\|_{2,1}^{-2} \geq s^{-1} T^{-1} C_\beta^{-2}$. Plugging back into (22), we have that

$$\begin{aligned} (I) &\geq 2^{-3} \phi_{\min}^2 \phi_{\max}^{-1} C_\beta^{-2} n s^{-1} T^{-1} \left(\min_{j \in \mathcal{M}^*} \sum_{t \in [T]} \beta_{t,j}^2 \right)^2 \\ &\geq 2^{-3} \phi_{\min}^2 \phi_{\max}^{-1} C_\beta^{-2} C_s^{-1} n^{1-\delta_s} T^{-1} \left(\min_{j \in \mathcal{M}^*} \sum_{t \in [T]} \beta_{t,j}^2 \right)^2 \end{aligned} \quad (24)$$

Next, we deal with the second term in (19). Recall that $\mathbf{X}_{t,j}^{(k)} = \mathbf{H}_{t,\mathcal{M}^{(k)}}^\perp \mathbf{X}_{t,j}$, so that $\|\mathbf{X}_{t,j}^{(k)}\|_2^2 \geq 2^{-1} n \phi_{\min}$, on the event \mathcal{E} . We have

$$\begin{aligned} &\sum_t \|\mathbf{H}_{t,j}^{(k)} (\mathbf{I}_{n \times n} - \mathbf{H}_{t,\mathcal{M}^{(k)}}) \boldsymbol{\epsilon}_t\|_2^2 \\ &= \sum_t \|\mathbf{X}_{t,j}^{(k)}\|^{-2} (\mathbf{X}_{t,j}' \mathbf{H}_{t,\mathcal{M}^{(k)}}^\perp \boldsymbol{\epsilon}_t)^2 \\ &\leq 2 \phi_{\min}^{-1} n^{-1} \max_{j \in \mathcal{M}_*} \max_{|\mathcal{M}| \leq m_{\max}^*} \sum_t (\mathbf{X}_{t,j}' \mathbf{H}_{t,\mathcal{M}}^\perp \boldsymbol{\epsilon}_t)^2. \end{aligned} \quad (25)$$

Under the conditions of the theorem, $\mathbf{X}_{t,j}' \mathbf{H}_{t,\mathcal{M}}^\perp \boldsymbol{\epsilon}_t$ is normally distributed with mean 0 and variance $\|\mathbf{H}_{t,\mathcal{M}}^\perp \mathbf{X}_{t,j}\|_2^2$. Furthermore,

$$\max_{j \in \mathcal{M}_*} \max_{|\mathcal{M}| \leq m_{\max}^*} \max_{t \in [T]} \|\mathbf{H}_{t,\mathcal{M}}^\perp \mathbf{X}_{t,j}\|_2^2 \leq 2n\phi_{\max}. \quad (26)$$

Plugging back in (25), we have

$$(II) \leq 2^2 \phi_{\min}^{-1} \phi_{\max} \max_{j \in \mathcal{M}_*} \max_{|\mathcal{M}| \leq m_{\max}^*} \chi_T^2, \quad (27)$$

where χ_T^2 denotes a chi-squared random variable with T degrees of freedom. The total number of possibilities for $j \in \mathcal{M}_*$ and $|\mathcal{M}| \leq m_{\max}^*$ is bounded by $p^{m_{\max}^*+2}$. Using Lemma 5, with $\epsilon = T(m_{\max}^* + 2) \log p$ and applying the union bound, we obtain

$$\begin{aligned} (II) &\leq 2^3 \phi_{\min}^{-1} \phi_{\max} T (m_{\max}^* + 2) \log p \\ &\leq 9 \phi_{\min}^{-1} \phi_{\max} C_p n^{\delta_p} T m_{\max}^* \end{aligned} \quad (28)$$

with probability at least

$$1 - p^{m_{\max}^*+2} \exp \left(-2T(m_{\max}^* + 2) \log(p) \left(1 - 2 \sqrt{\frac{1}{2(m_{\max}^* + 2) \log(p)}} \right) \right). \quad (29)$$

Going back to (19), we have the following

$$\begin{aligned}
n^{-1}T^{-1}\Delta(k) &\geq 2^{-3}\phi_{\min}^2\phi_{\max}^{-1}C_{\beta}^{-2}C_s^{-1}n^{-\delta_s}T^{-2}\left(\min_{j\in\mathcal{M}^*}\sum_{t\in[T]}\beta_{t,j}^2\right)^2 \\
&\quad - 9\phi_{\min}^{-1}\phi_{\max}C_p n^{\delta_p-1}m_{\max}^* \\
&\geq 2^{-3}\phi_{\min}^2\phi_{\max}^{-1}C_{\beta}^{-2}C_s^{-1}c_{\beta}^2n^{-\delta_s-2\delta_{\min}} \\
&\quad - 9\phi_{\min}^{-1}\phi_{\max}C_p n^{\delta_p-1}m_{\max}^* \\
&\geq 2^{-3}\phi_{\min}^2\phi_{\max}^{-1}C_{\beta}^{-2}C_s^{-1}c_{\beta}^2n^{-\delta_s-2\delta_{\min}} \\
&\quad \times (1 - 72\phi_{\min}^{-3}\phi_{\max}^2C_{\beta}^2C_pC_sc_{\beta}^{-2}n^{\delta_s+2\delta_{\min}+\delta_p-1}m_{\max}^*).
\end{aligned} \tag{30}$$

Since the bound in (30) holds uniformly for $k \in \{1, \dots, m_{\text{one}}^*\}$, we have that $n^{-1}T^{-1}\sum_{t\in[T]}\|\mathbf{y}_t\|_2^2 \geq n^{-1}T^{-1}\sum_{k=1}^{m_{\text{one}}^*}\Delta(k)$. Setting

$$m_{\text{one}}^* = \lfloor 2^4\phi_{\min}^{-2}\phi_{\max}C_{\beta}^2C_sc_{\beta}^{-2}n^{\delta_s+2\delta_{\min}} \rfloor \tag{31}$$

and recalling that $m_{\max}^* = sm_{\text{one}}^*$, the lower bound becomes

$$n^{-1}T^{-1}\sum_{t\in[T]}\|\mathbf{y}_t\|_2^2 \geq 2(1 - Cn^{3\delta_s+4\delta_{\min}+\delta_p-1}), \tag{32}$$

for a positive constant C independent of p, n, s and T . Under the conditions of the theorem, the right side of (32) is bounded below by 2. We have arrived at a contradiction, since under the assumptions $\text{Var}(y_{t,i}) = 1$ and by the weak law of large numbers, $n^{-1}T^{-1}\sum_{t\in[T]}\|\mathbf{y}_t\|_2^2 \rightarrow 1$ in probability. Therefore, at least one relevant variable will be selected in m_{one}^* steps.

To complete the proof, we lower bound the probability in (28) and the probability of the event \mathcal{E} . Plugging in the value for m_{\max}^* , the probability in (28) can be lower bounded by $1 - \exp(-C(2T-1)n^{2\delta_s+2\delta_{\min}+\delta_p})$ for some positive constant C . The probability of the event \mathcal{E} is lower bounded, using Lemma 3 together with the union bound, as $1 - C_1 \exp(-C_2 \frac{n^{1-6\delta_s-6\delta_{\min}}}{\max\{\log p, \log T\}})$, for some positive constants C_1 and C_2 . Both of these probabilities converge to 1 under the conditions of the theorem.

6.2 Proof of Theorem 2

To prove the theorem, we use the same strategy as in Wang (2009). From Theorem 1, we have that $\mathbb{P}[\exists k \in \{0, \dots, n-1\} : \mathcal{M}_* \subseteq \mathcal{M}^{(k)}] \rightarrow 1$, so $k_{\min} := \min_{k \in \{0, \dots, n-1\}} \{k : \mathcal{M}_* \subseteq \mathcal{M}^{(k)}\}$ is well defined and $k_{\min} \leq m_{\max}^*$, for m_{\max}^* defined in (13). We show that

$$\mathbb{P}[\min_{k \in \{0, \dots, k_{\min}-1\}} (\text{BIC}(\mathcal{M}^{(k)}) - \text{BIC}(\mathcal{M}^{(k+1)})) > 0] \rightarrow 1, \tag{33}$$

so that $\mathbb{P}[\hat{s} < k_{\min}] \rightarrow 0$ as $n \rightarrow \infty$. We proceed by lower bounding the difference in the BIC scores as

$$\begin{aligned} \text{BIC}(\mathcal{M}^{(k)}) - \text{BIC}(\mathcal{M}^{(k+1)}) &= \log \left(\frac{\text{RSS}(\mathcal{M}^{(k)})}{\text{RSS}(\mathcal{M}^{(k+1)})} \right) - \frac{\log(n) + 2 \log(p)}{n} \\ &\geq \log \left(1 + \frac{\text{RSS}(\mathcal{M}^{(k)}) - \text{RSS}(\mathcal{M}^{(k+1)})}{\text{RSS}(\mathcal{M}^{(k+1)})} \right) - 3n^{-1} \log(p), \end{aligned} \quad (34)$$

where we have assumed $p > n$. Define the event $\mathcal{A} := \{n^{-1}T^{-1} \sum_{t \in [T]} \|\mathbf{y}_t\|_2^2 \leq 2\}$. Note that $\text{RSS}(\mathcal{M}^{(k+1)}) \leq \sum_{t \in [T]} \|\mathbf{y}_t\|_2^2$, so on the event \mathcal{A} the difference in the BIC scores is lower bounded as

$$\log(1 + 2n^{-1}T^{-1}\Delta(k)) - 3n^{-1} \log(p), \quad (35)$$

where $\Delta(k)$ is defined in (17). Using the fact that $\log(1+x) \geq \min(\log(2), 2^{-1}x)$ and the lower bound from (30), we have

$$\text{BIC}(\mathcal{M}^{(k)}) - \text{BIC}(\mathcal{M}^{(k+1)}) \geq \min(\log 2, Cn^{-\delta_s - 2\delta_{\min}}) - 3n^{-1} \log p, \quad (36)$$

for some positive constant C . It is easy to check that $\log 2 - 3n^{-1} \log p > 0$ and $Cn^{-\delta_s - 2\delta_{\min}} - 3n^{-1} \log p > 0$ under the conditions of the theorem. The lower bound in (36) is uniform for $k \in \{0, \dots, k_{\min}\}$, so the proof is complete if we show that $\mathbb{P}[\mathcal{A}] \rightarrow 1$. But this easily follows from the tail bounds on the central chi-squared random variable given in Lemma 4.

6.3 Collection of known results

In what follows, C_1, C_2, \dots will denote arbitrary positive constants.

The following result on the minimum eigenvalue of sub-matrices of the covariance matrix $\hat{\Sigma}$ is quite standard (e.g. Zhou et al. (2009), Wang (2009) or Bickel et al. (2009)).

Lemma 3. *Let $\mathbf{x} \sim \mathcal{N}(0, \Sigma)$ and $\hat{\Sigma} = n^{-1} \sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i'$ be the empirical estimate from n independent realizations of \mathbf{x} . Denote $\Sigma = [\sigma_{ab}]$ and $\hat{\Sigma} = [\hat{\sigma}_{ab}]$. Assume $\phi_{\min} \leq \Lambda_{\min}(\Sigma) \leq \Lambda_{\max}(\Sigma) \leq \phi_{\max}$. Then*

$$\mathbb{P}[\max_{\mathcal{M} \subseteq [p], |\mathcal{M}| < s} \Lambda_{\max}(\hat{\Sigma}_{\mathcal{M}}) \geq 2\phi_{\max}] \leq C_1 \exp(-C_2 \frac{n}{s^2} + s \log p) \quad (37)$$

and

$$\mathbb{P}[\min_{\mathcal{M} \subseteq [p], |\mathcal{M}| < s} \Lambda_{\min}(\hat{\Sigma}_{\mathcal{M}}) \leq \phi_{\min}/2] \leq C_3 \exp(-C_4 \frac{n}{s^2} + s \log p). \quad (38)$$

The following tail bounds for the chi-squared distribution are taken from Laurent and Massart (2000).

Lemma 4. *Let χ_n^2 be a central chi-squared r.v. with n degrees of freedom. For any positive ϵ ,*

$$\mathbb{P}[\chi_n^2 \geq n + 2\sqrt{n\epsilon} + 2\epsilon] \leq \exp(-\epsilon) \quad (39)$$

$$\mathbb{P}[\chi_n^2 \leq n - 2\sqrt{n\epsilon}] \leq \exp(-\epsilon). \quad (40)$$

We also make use of the result taken from Obozinski et al. (2009), which bounds the maximum of a collection of chi-squared random variables.

Lemma 5. *Let X_1, \dots, X_m be i.i.d. central chi-squared r.v. with n degrees of freedom. Then for any $\epsilon > n$,*

$$\mathbb{P}[\max_{i \in [m]} X_i \geq 2\epsilon] \leq m \exp(-\epsilon(1 - 2\sqrt{\frac{n}{\epsilon}})). \quad (41)$$

6.4 Tables with simulation results

Simulation 1: $(n, p, s, T) = (500, 20000, 18, 500)$, $T_{\text{non-zero}} = 500$

Method name		$\mathcal{M}_* \subseteq \hat{S}$	Correct zeros	Incorrect zeros	$\mathcal{M}_* = \hat{S}$	$ \hat{S} $	$\ \mathbf{B} - \hat{\mathbf{B}}\ _2^2$	R^2
SNR = 15								
Union Support	SIS-ALASSO	100.0	100.0	0.0	10.0	20.2	-	-
	ISIS-ALASSO	100.0	100.0	0.0	18.0	19.6	-	-
	OMP	100.0	100.0	0.0	0.0	23.9	-	-
	S-OMP	100.0	100.0	0.0	100.0	18.0	-	-
	S-OMP-ALASSO	100.0	100.0	0.0	100.0	18.0	-	-
Exact Support	SIS-ALASSO	0.0	100.0	0.7	0.0	8940.5	0.97	0.93
	ISIS-ALASSO	100.0	100.0	0.0	18.0	9001.6	0.33	0.93
	OMP	100.0	100.0	0.0	0.0	9005.9	0.20	0.93
	S-OMP-ALASSO	100.0	100.0	0.0	100.0	9000.0	0.20	0.93
SNR = 10								
Union Support	SIS-ALASSO	100.0	100.0	0.0	0.0	25.3	-	-
	ISIS-ALASSO	100.0	100.0	0.0	0.0	25.7	-	-
	OMP	100.0	100.0	0.0	0.0	23.9	-	-
	S-OMP	100.0	100.0	0.0	100.0	18.0	-	-
	S-OMP-ALASSO	100.0	100.0	0.0	100.0	18.0	-	-
Exact Support	SIS-ALASSO	0.0	100.0	1.6	0.0	8861.0	2.06	0.89
	ISIS-ALASSO	100.0	100.0	0.0	0.0	9007.7	0.65	0.90
	OMP	100.0	100.0	0.0	0.0	9005.9	0.31	0.91
	S-OMP-ALASSO	65.0	100.0	0.1	65.0	8987.4	0.41	0.90
SNR = 5								
Union Support	SIS-ALASSO	100.0	100.0	0.0	64.0	18.4	-	-
	ISIS-ALASSO	100.0	100.0	0.0	57.0	18.6	-	-
	OMP	100.0	100.0	0.0	0.0	24.0	-	-
	S-OMP	100.0	100.0	0.0	100.0	18.0	-	-
	S-OMP-ALASSO	100.0	100.0	0.0	100.0	18.0	-	-
Exact Support	SIS-ALASSO	0.0	100.0	92.8	0.0	645.8	74.61	0.06
	ISIS-ALASSO	0.0	100.0	90.9	0.0	822.2	73.06	0.07
	OMP	100.0	100.0	0.0	0.0	9006.0	0.61	0.83
	S-OMP-ALASSO	0.0	100.0	70.3	0.0	2668.9	56.65	0.24
SNR = 1								
Union Support	SIS-ALASSO	0.0	100.0	99.9	0.0	0.0	-	-
	ISIS-ALASSO	0.0	100.0	100.0	0.0	0.0	-	-
	OMP	100.0	100.0	0.0	0.0	25.9	-	-
	S-OMP	0.0	100.0	94.4	0.0	1.0	-	-
	S-OMP-ALASSO	0.0	100.0	99.0	0.0	0.2	-	-
Exact Support	SIS-ALASSO	0.0	100.0	100.0	0.0	0.0	80.27	-0.00
	ISIS-ALASSO	0.0	100.0	100.0	0.0	0.0	80.27	-0.00
	OMP	0.0	100.0	86.5	0.0	1222.8	71.40	0.05
	S-OMP-ALASSO	0.0	100.0	100.0	0.0	0.2	80.27	-0.00

Simulation 1: $(n, p, s, T) = (500, 20000, 18, 500)$, $T_{\text{non-zero}} = 300$

Method name		$\mathcal{M}_* \subseteq \hat{S}$	Correct zeros	Incorrect zeros	$\mathcal{M}_* = \hat{S}$	$ \hat{S} $	$\ \mathbf{B} - \hat{\mathbf{B}}\ _2^2$	R^2
SNR = 15								
Union Support	SIS-ALASSO	100.0	100.0	0.0	97.0	18.0	-	-
	ISIS-ALASSO	100.0	100.0	0.0	98.0	18.0	-	-
	OMP	100.0	100.0	0.0	0.0	23.0	-	-
	S-OMP	100.0	100.0	0.0	100.0	18.0	-	-
	S-OMP-ALASSO	100.0	100.0	0.0	100.0	18.0	-	-
Exact Support	SIS-ALASSO	55.0	100.0	0.0	53.0	5399.3	0.10	0.93
	ISIS-ALASSO	100.0	100.0	0.0	98.0	5400.0	0.09	0.93
	OMP	100.0	100.0	0.0	0.0	5405.0	0.07	0.93
	S-OMP-ALASSO	100.0	100.0	0.0	100.0	5400.0	0.07	0.93
SNR = 10								
Union Support	SIS-ALASSO	100.0	100.0	0.0	82.0	18.2	-	-
	ISIS-ALASSO	100.0	100.0	0.0	91.0	18.1	-	-
	OMP	100.0	100.0	0.0	0.0	23.0	-	-
	S-OMP	100.0	100.0	0.0	100.0	18.0	-	-
	S-OMP-ALASSO	100.0	100.0	0.0	100.0	18.0	-	-
Exact Support	SIS-ALASSO	42.0	100.0	0.0	33.0	5399.2	0.18	0.90
	ISIS-ALASSO	100.0	100.0	0.0	91.0	5400.1	0.16	0.90
	OMP	100.0	100.0	0.0	0.0	5405.0	0.11	0.90
	S-OMP-ALASSO	100.0	100.0	0.0	100.0	5400.0	0.11	0.90
SNR = 5								
Union Support	SIS-ALASSO	100.0	100.0	0.0	3.0	21.1	-	-
	ISIS-ALASSO	100.0	100.0	0.0	6.0	20.8	-	-
	OMP	100.0	100.0	0.0	0.0	23.0	-	-
	S-OMP	100.0	100.0	0.0	100.0	18.0	-	-
	S-OMP-ALASSO	100.0	100.0	0.0	100.0	18.0	-	-
Exact Support	SIS-ALASSO	24.0	100.0	0.0	1.0	5400.9	0.61	0.82
	ISIS-ALASSO	99.0	100.0	0.0	6.0	5402.8	0.52	0.82
	OMP	100.0	100.0	0.0	0.0	5405.0	0.22	0.82
	S-OMP-ALASSO	100.0	100.0	0.0	100.0	5400.0	0.23	0.82
SNR = 1								
Union Support	SIS-ALASSO	0.0	100.0	97.9	0.0	0.4	-	-
	ISIS-ALASSO	0.0	100.0	97.9	0.0	0.4	-	-
	OMP	100.0	100.0	0.0	0.0	25.9	-	-
	S-OMP	0.0	100.0	94.4	0.0	1.0	-	-
	S-OMP-ALASSO	0.0	100.0	94.4	0.0	1.0	-	-
Exact Support	SIS-ALASSO	0.0	100.0	100.0	0.0	0.4	48.16	-0.00
	ISIS-ALASSO	0.0	100.0	100.0	0.0	0.4	48.16	-0.00
	OMP	0.0	100.0	10.2	0.0	4858.1	5.76	0.43
	S-OMP-ALASSO	0.0	100.0	99.9	0.0	6.1	48.12	-0.00

Simulation 1: $(n, p, s, T) = (500, 20000, 18, 500)$, $T_{\text{non-zero}} = 100$

Method name		$\mathcal{M}_* \subseteq \hat{S}$	Correct zeros	Incorrect zeros	$\mathcal{M}_* = \hat{S}$	$ \hat{S} $	$\ \mathbf{B} - \hat{\mathbf{B}}\ _2^2$	R^2
SNR = 15								
Union Support	SIS-ALASSO	100.0	100.0	0.0	100.0	18.0	-	-
	ISIS-ALASSO	100.0	100.0	0.0	100.0	18.0	-	-
	OMP	100.0	99.9	0.0	0.0	28.8	-	-
	S-OMP	100.0	100.0	0.0	100.0	18.0	-	-
	S-OMP-ALASSO	100.0	100.0	0.0	100.0	18.0	-	-
Exact Support	SIS-ALASSO	100.0	100.0	0.0	100.0	1800.0	0.01	0.91
	ISIS-ALASSO	100.0	100.0	0.0	100.0	1800.0	0.01	0.91
	OMP	100.0	100.0	0.0	0.0	1810.8	0.01	0.91
	S-OMP-ALASSO	100.0	100.0	0.0	100.0	1800.0	0.01	0.91
SNR = 10								
Union Support	SIS-ALASSO	100.0	100.0	0.0	100.0	18.0	-	-
	ISIS-ALASSO	100.0	100.0	0.0	100.0	18.0	-	-
	OMP	100.0	99.9	0.0	0.0	28.8	-	-
	S-OMP	100.0	100.0	0.0	100.0	18.0	-	-
	S-OMP-ALASSO	100.0	100.0	0.0	100.0	18.0	-	-
Exact Support	SIS-ALASSO	100.0	100.0	0.0	100.0	1800.0	0.01	0.88
	ISIS-ALASSO	100.0	100.0	0.0	100.0	1800.0	0.01	0.88
	OMP	100.0	100.0	0.0	0.0	1810.8	0.01	0.88
	S-OMP-ALASSO	100.0	100.0	0.0	100.0	1800.0	0.01	0.88
SNR = 5								
Union Support	SIS-ALASSO	100.0	100.0	0.0	100.0	18.0	-	-
	ISIS-ALASSO	100.0	100.0	0.0	100.0	18.0	-	-
	OMP	100.0	99.9	0.0	0.0	28.8	-	-
	S-OMP	100.0	100.0	0.0	100.0	18.0	-	-
	S-OMP-ALASSO	100.0	100.0	0.0	100.0	18.0	-	-
Exact Support	SIS-ALASSO	100.0	100.0	0.0	100.0	1800.0	0.04	0.79
	ISIS-ALASSO	100.0	100.0	0.0	100.0	1800.0	0.03	0.79
	OMP	100.0	100.0	0.0	0.0	1810.8	0.03	0.79
	S-OMP-ALASSO	100.0	100.0	0.0	100.0	1800.0	0.02	0.79
SNR = 1								
Union Support	SIS-ALASSO	100.0	100.0	0.0	19.0	19.6	-	-
	ISIS-ALASSO	100.0	100.0	0.0	35.0	19.0	-	-
	OMP	100.0	99.9	0.0	0.0	28.8	-	-
	S-OMP	0.0	100.0	94.4	0.0	1.0	-	-
	S-OMP-ALASSO	0.0	100.0	94.4	0.0	1.0	-	-
Exact Support	SIS-ALASSO	59.0	100.0	0.0	10.0	1800.9	0.74	0.45
	ISIS-ALASSO	89.0	100.0	0.0	32.0	1800.8	0.63	0.45
	OMP	100.0	100.0	0.0	0.0	1810.8	0.13	0.47
	S-OMP-ALASSO	0.0	100.0	95.3	0.0	84.6	15.31	0.02

Simulation 2.a: $(n, p, s, T) = (200, 5000, 10, 500)$, $T_{\text{non-zero}} = 400$

Method name		$\mathcal{M}_* \subseteq \hat{S}$	Correct zeros	Incorrect zeros	$\mathcal{M}_* = \hat{S}$	$ \hat{S} $	$\ \mathbf{B} - \hat{\mathbf{B}}\ _2^2$	R^2
SNR = 10								
Union Support	SIS-ALASSO	100.0	100.0	0.0	39.0	10.9	-	-
	ISIS-ALASSO	100.0	100.0	0.0	12.0	12.2	-	-
	OMP	100.0	99.8	0.0	0.0	21.6	-	-
	S-OMP	100.0	100.0	0.0	100.0	10.0	-	-
	S-OMP-ALASSO	100.0	100.0	0.0	100.0	10.0	-	-
Exact Support	SIS-ALASSO	0.0	100.0	6.4	0.0	3746.6	3.58	0.85
	ISIS-ALASSO	41.0	100.0	0.2	3.0	3992.8	0.53	0.90
	OMP	100.0	100.0	0.0	0.0	4011.7	0.22	0.90
	S-OMP-ALASSO	99.0	100.0	0.0	98.0	3999.9	0.22	0.90
SNR = 5								
Union Support	SIS-ALASSO	100.0	100.0	0.0	45.0	11.0	-	-
	ISIS-ALASSO	100.0	100.0	0.0	37.0	10.9	-	-
	OMP	100.0	99.8	0.0	0.0	22.2	-	-
	S-OMP	100.0	100.0	0.0	100.0	10.0	-	-
	S-OMP-ALASSO	100.0	100.0	0.0	100.0	10.0	-	-
Exact Support	SIS-ALASSO	0.0	100.0	65.9	0.0	1363.5	33.32	0.30
	ISIS-ALASSO	0.0	100.0	63.1	0.0	1477.0	31.89	0.33
	OMP	100.0	100.0	0.0	0.0	4012.2	0.45	0.82
	S-OMP-ALASSO	0.0	100.0	48.0	0.0	2081.5	24.19	0.46
SNR = 1								
Union Support	SIS-ALASSO	0.0	100.0	98.2	0.0	0.2	-	-
	ISIS-ALASSO	0.0	100.0	98.7	0.0	0.1	-	-
	OMP	100.0	99.5	0.0	0.0	35.2	-	-
	S-OMP	0.0	100.0	90.0	0.0	1.0	-	-
	S-OMP-ALASSO	0.0	100.0	95.4	0.0	0.5	-	-
Exact Support	SIS-ALASSO	0.0	100.0	100.0	0.0	0.2	49.94	-0.00
	ISIS-ALASSO	0.0	100.0	100.0	0.0	0.1	49.94	-0.00
	OMP	0.0	100.0	76.5	0.0	964.4	40.05	0.09
	S-OMP-ALASSO	0.0	100.0	100.0	0.0	0.8	49.94	-0.00

Simulation 2.a: $(n, p, s, T) = (200, 5000, 10, 500)$, $T_{\text{non-zero}} = 250$

Method name		$\mathcal{M}_* \subseteq \hat{S}$	Correct zeros	Incorrect zeros	$\mathcal{M}_* = \hat{S}$	$ \hat{S} $	$\ \mathbf{B} - \hat{\mathbf{B}}\ _2^2$	R^2
SNR = 10								
Union Support	SIS-ALASSO	100.0	100.0	0.0	99.0	10.0	-	-
	ISIS-ALASSO	100.0	100.0	0.0	98.0	10.0	-	-
	OMP	100.0	99.8	0.0	0.0	19.9	-	-
	S-OMP	100.0	100.0	0.0	100.0	10.0	-	-
	S-OMP-ALASSO	100.0	100.0	0.0	100.0	10.0	-	-
Exact Support	SIS-ALASSO	22.0	100.0	0.2	22.0	2495.4	0.19	0.89
	ISIS-ALASSO	100.0	100.0	0.0	98.0	2500.0	0.12	0.89
	OMP	100.0	100.0	0.0	0.0	2509.9	0.09	0.90
	S-OMP-ALASSO	100.0	100.0	0.0	100.0	2500.0	0.08	0.90
SNR = 5								
Union Support	SIS-ALASSO	100.0	100.0	0.0	44.0	10.8	-	-
	ISIS-ALASSO	100.0	100.0	0.0	46.0	10.8	-	-
	OMP	100.0	99.8	0.0	0.0	19.9	-	-
	S-OMP	100.0	100.0	0.0	100.0	10.0	-	-
	S-OMP-ALASSO	100.0	100.0	0.0	100.0	10.0	-	-
Exact Support	SIS-ALASSO	12.0	100.0	0.9	6.0	2479.5	0.69	0.80
	ISIS-ALASSO	62.0	100.0	0.2	29.0	2496.7	0.43	0.81
	OMP	100.0	100.0	0.0	0.0	2509.9	0.18	0.81
	S-OMP-ALASSO	95.0	100.0	0.0	95.0	2499.6	0.18	0.81
SNR = 1								
Union Support	SIS-ALASSO	0.0	100.0	65.3	0.0	3.5	-	-
	ISIS-ALASSO	0.0	100.0	61.3	0.0	3.9	-	-
	OMP	100.0	99.7	0.0	0.0	24.7	-	-
	S-OMP	0.0	100.0	90.0	0.0	1.0	-	-
	S-OMP-ALASSO	0.0	100.0	90.0	0.0	1.0	-	-
Exact Support	SIS-ALASSO	0.0	100.0	99.8	0.0	4.6	31.16	-0.00
	ISIS-ALASSO	0.0	100.0	99.8	0.0	5.2	31.15	-0.00
	OMP	0.0	100.0	17.2	0.0	2083.7	6.09	0.39
	S-OMP-ALASSO	0.0	100.0	99.6	0.0	10.4	31.11	-0.00

Simulation 2.a: $(n, p, s, T) = (200, 5000, 10, 500)$, $T_{\text{non-zero}} = 100$

Method name		$\mathcal{M}_* \subseteq \hat{S}$	Correct zeros	Incorrect zeros	$\mathcal{M}_* = \hat{S}$	$ \hat{S} $	$\ \mathbf{B} - \hat{\mathbf{B}}\ _2^2$	R^2
SNR = 10								
Union Support	SIS-ALASSO	100.0	100.0	0.0	100.0	10.0	-	-
	ISIS-ALASSO	100.0	100.0	0.0	100.0	10.0	-	-
	OMP	100.0	98.8	0.0	0.0	69.8	-	-
	S-OMP	100.0	100.0	0.0	100.0	10.0	-	-
	S-OMP-ALASSO	100.0	100.0	0.0	100.0	10.0	-	-
Exact Support	SIS-ALASSO	98.0	100.0	0.0	98.0	1000.0	0.02	0.80
	ISIS-ALASSO	100.0	100.0	0.0	100.0	1000.0	0.01	0.80
	OMP	100.0	100.0	0.0	0.0	1060.2	0.02	0.79
	S-OMP-ALASSO	100.0	100.0	0.0	100.0	1000.0	0.01	0.80
SNR = 5								
Union Support	SIS-ALASSO	100.0	100.0	0.0	100.0	10.0	-	-
	ISIS-ALASSO	100.0	100.0	0.0	100.0	10.0	-	-
	OMP	100.0	98.8	0.0	0.0	69.8	-	-
	S-OMP	100.0	100.0	0.0	100.0	10.0	-	-
	S-OMP-ALASSO	100.0	100.0	0.0	100.0	10.0	-	-
Exact Support	SIS-ALASSO	98.0	100.0	0.0	98.0	1000.0	0.04	0.73
	ISIS-ALASSO	100.0	100.0	0.0	100.0	1000.0	0.04	0.73
	OMP	100.0	100.0	0.0	0.0	1060.2	0.05	0.72
	S-OMP-ALASSO	100.0	100.0	0.0	100.0	1000.0	0.03	0.73
SNR = 1								
Union Support	SIS-ALASSO	100.0	100.0	0.0	61.0	10.6	-	-
	ISIS-ALASSO	100.0	100.0	0.0	60.0	10.5	-	-
	OMP	100.0	98.8	0.0	0.0	69.8	-	-
	S-OMP	0.0	100.0	90.0	0.0	1.0	-	-
	S-OMP-ALASSO	0.0	100.0	90.0	0.0	1.0	-	-
Exact Support	SIS-ALASSO	0.0	100.0	12.7	0.0	873.9	2.23	0.37
	ISIS-ALASSO	0.0	100.0	9.8	0.0	902.8	1.79	0.38
	OMP	100.0	100.0	0.0	0.0	1060.2	0.25	0.42
	S-OMP-ALASSO	0.0	100.0	93.3	0.0	67.4	11.66	0.03

Simulation 2.b: $(n, p, s, T) = (200, 5000, 10, 750)$, $T_{\text{non-zero}} = 600$

Method name		$\mathcal{M}_* \subseteq \hat{S}$	Correct zeros	Incorrect zeros	$\mathcal{M}_* = \hat{S}$	$ \hat{S} $	$\ \mathbf{B} - \hat{\mathbf{B}}\ _2^2$	R^2
SNR = 10								
Union Support	SIS-ALASSO	100.0	100.0	0.0	25.0	11.3	-	-
	ISIS-ALASSO	100.0	99.9	0.0	5.0	13.3	-	-
	OMP	100.0	99.7	0.0	0.0	26.6	-	-
	S-OMP	100.0	100.0	0.0	100.0	10.0	-	-
	S-OMP-ALASSO	100.0	100.0	0.0	100.0	10.0	-	-
Exact Support	SIS-ALASSO	0.0	100.0	6.9	0.0	5585.0	3.87	0.84
	ISIS-ALASSO	29.0	100.0	0.3	4.0	5986.6	0.56	0.90
	OMP	100.0	100.0	0.0	0.0	6016.7	0.22	0.90
	S-OMP-ALASSO	91.0	100.0	0.0	91.0	5999.1	0.23	0.90
SNR = 5								
Union Support	SIS-ALASSO	100.0	100.0	0.0	27.0	11.4	-	-
	ISIS-ALASSO	100.0	100.0	0.0	28.0	11.3	-	-
	OMP	100.0	99.7	0.0	0.0	27.3	-	-
	S-OMP	100.0	100.0	0.0	100.0	10.0	-	-
	S-OMP-ALASSO	100.0	100.0	0.0	100.0	10.0	-	-
Exact Support	SIS-ALASSO	0.0	100.0	66.5	0.0	2011.9	33.60	0.30
	ISIS-ALASSO	0.0	100.0	63.6	0.0	2185.7	32.14	0.32
	OMP	100.0	100.0	0.0	0.0	6017.5	0.45	0.82
	S-OMP-ALASSO	0.0	100.0	48.3	0.0	3104.4	24.34	0.45
SNR = 1								
Union Support	SIS-ALASSO	0.0	100.0	97.8	0.0	0.2	-	-
	ISIS-ALASSO	0.0	100.0	98.2	0.0	0.2	-	-
	OMP	100.0	99.2	0.0	0.0	47.6	-	-
	S-OMP	0.0	100.0	90.0	0.0	1.0	-	-
	S-OMP-ALASSO	0.0	100.0	94.7	0.0	0.5	-	-
Exact Support	SIS-ALASSO	0.0	100.0	100.0	0.0	0.2	49.94	-0.01
	ISIS-ALASSO	0.0	100.0	100.0	0.0	0.2	49.94	-0.01
	OMP	0.0	100.0	76.7	0.0	1436.7	40.13	0.09
	S-OMP-ALASSO	0.0	100.0	100.0	0.0	1.0	49.94	-0.01

Simulation 2.b: $(n, p, s, T) = (200, 5000, 10, 750)$, $T_{\text{non-zero}} = 375$

Method name		$\mathcal{M}_* \subseteq \hat{S}$	Correct zeros	Incorrect zeros	$\mathcal{M}_* = \hat{S}$	$ \hat{S} $	$\ \mathbf{B} - \hat{\mathbf{B}}\ _2^2$	R^2
SNR = 10								
Union Support	SIS-ALASSO	100.0	100.0	0.0	99.0	10.0	-	-
	ISIS-ALASSO	100.0	100.0	0.0	93.0	10.1	-	-
	OMP	100.0	99.7	0.0	0.0	24.7	-	-
	S-OMP	100.0	100.0	0.0	100.0	10.0	-	-
	S-OMP-ALASSO	100.0	100.0	0.0	100.0	10.0	-	-
Exact Support	SIS-ALASSO	16.0	100.0	0.2	16.0	3741.3	0.21	0.89
	ISIS-ALASSO	100.0	100.0	0.0	93.0	3750.1	0.12	0.89
	OMP	100.0	100.0	0.0	0.0	3764.8	0.09	0.89
	S-OMP-ALASSO	100.0	100.0	0.0	100.0	3750.0	0.09	0.89
SNR = 5								
Union Support	SIS-ALASSO	100.0	100.0	0.0	41.0	10.9	-	-
	ISIS-ALASSO	100.0	100.0	0.0	25.0	11.4	-	-
	OMP	100.0	99.7	0.0	0.0	24.7	-	-
	S-OMP	100.0	100.0	0.0	100.0	10.0	-	-
	S-OMP-ALASSO	100.0	100.0	0.0	100.0	10.0	-	-
Exact Support	SIS-ALASSO	6.0	100.0	1.0	3.0	3713.5	0.73	0.80
	ISIS-ALASSO	53.0	100.0	0.2	13.0	3744.9	0.43	0.80
	OMP	100.0	100.0	0.0	0.0	3764.8	0.18	0.81
	S-OMP-ALASSO	91.0	100.0	0.0	91.0	3749.0	0.19	0.81
SNR = 1								
Union Support	SIS-ALASSO	0.0	100.0	55.8	0.0	4.4	-	-
	ISIS-ALASSO	1.0	100.0	52.8	1.0	4.7	-	-
	OMP	100.0	99.6	0.0	0.0	32.0	-	-
	S-OMP	0.0	100.0	90.0	0.0	1.0	-	-
	S-OMP-ALASSO	0.0	100.0	90.0	0.0	1.0	-	-
Exact Support	SIS-ALASSO	0.0	100.0	99.8	0.0	6.6	31.16	-0.00
	ISIS-ALASSO	0.0	100.0	99.8	0.0	7.3	31.16	-0.00
	OMP	0.0	100.0	17.6	0.0	3111.8	6.21	0.39
	S-OMP-ALASSO	0.0	100.0	99.6	0.0	15.1	31.11	-0.00

Simulation 2.b: $(n, p, s, T) = (200, 5000, 10, 750)$, $T_{\text{non-zero}} = 150$

Method name		$\mathcal{M}_* \subseteq \hat{S}$	Correct zeros	Incorrect zeros	$\mathcal{M}_* = \hat{S}$	$ \hat{S} $	$\ \mathbf{B} - \hat{\mathbf{B}}\ _2^2$	R^2
SNR = 10								
Union Support	SIS-ALASSO	100.0	100.0	0.0	100.0	10.0	-	-
	ISIS-ALASSO	100.0	100.0	0.0	100.0	10.0	-	-
	OMP	100.0	98.0	0.0	0.0	108.5	-	-
	S-OMP	100.0	100.0	0.0	100.0	10.0	-	-
	S-OMP-ALASSO	100.0	100.0	0.0	100.0	10.0	-	-
Exact Support	SIS-ALASSO	98.0	100.0	0.0	98.0	1500.0	0.02	0.79
	ISIS-ALASSO	100.0	100.0	0.0	100.0	1500.0	0.02	0.79
	OMP	100.0	100.0	0.0	0.0	1599.5	0.03	0.78
	S-OMP-ALASSO	100.0	100.0	0.0	100.0	1500.0	0.01	0.79
SNR = 5								
Union Support	SIS-ALASSO	100.0	100.0	0.0	100.0	10.0	-	-
	ISIS-ALASSO	100.0	100.0	0.0	100.0	10.0	-	-
	OMP	100.0	98.0	0.0	0.0	108.5	-	-
	S-OMP	100.0	100.0	0.0	100.0	10.0	-	-
	S-OMP-ALASSO	100.0	100.0	0.0	100.0	10.0	-	-
Exact Support	SIS-ALASSO	98.0	100.0	0.0	98.0	1500.0	0.04	0.72
	ISIS-ALASSO	100.0	100.0	0.0	100.0	1500.0	0.04	0.72
	OMP	100.0	100.0	0.0	0.0	1599.5	0.05	0.71
	S-OMP-ALASSO	100.0	100.0	0.0	100.0	1500.0	0.03	0.72
SNR = 1								
Union Support	SIS-ALASSO	100.0	100.0	0.0	46.0	10.8	-	-
	ISIS-ALASSO	100.0	100.0	0.0	42.0	10.8	-	-
	OMP	100.0	98.0	0.0	0.0	108.5	-	-
	S-OMP	0.0	100.0	90.0	0.0	1.0	-	-
	S-OMP-ALASSO	0.0	100.0	90.0	0.0	1.0	-	-
Exact Support	SIS-ALASSO	0.0	100.0	12.1	0.0	1318.9	2.16	0.37
	ISIS-ALASSO	0.0	100.0	9.4	0.0	1360.3	1.74	0.38
	OMP	100.0	100.0	0.0	0.0	1599.5	0.26	0.42
	S-OMP-ALASSO	0.0	100.0	93.4	0.0	98.9	11.68	0.03

Simulation 2.c: $(n, p, s, T) = (200, 5000, 10, 1000)$, $T_{\text{non-zero}} = 800$

Method name		$\mathcal{M}_* \subseteq \hat{S}$	Correct zeros	Incorrect zeros	$\mathcal{M}_* = \hat{S}$	$ \hat{S} $	$\ \mathbf{B} - \hat{\mathbf{B}}\ _2^2$	R^2
SNR = 10								
Union Support	SIS-ALASSO	100.0	100.0	0.0	21.0	11.7	-	-
	ISIS-ALASSO	100.0	99.9	0.0	5.0	14.4	-	-
	OMP	100.0	99.6	0.0	0.0	32.0	-	-
	S-OMP	100.0	100.0	0.0	100.0	10.0	-	-
	S-OMP-ALASSO	100.0	100.0	0.0	100.0	10.0	-	-
Exact Support	SIS-ALASSO	0.0	100.0	7.7	0.0	7382.7	4.26	0.84
	ISIS-ALASSO	17.0	100.0	0.4	1.0	7976.0	0.60	0.90
	OMP	100.0	100.0	0.0	0.0	8022.1	0.22	0.90
	S-OMP-ALASSO	86.0	100.0	0.0	86.0	7998.3	0.23	0.90
SNR = 5								
Union Support	SIS-ALASSO	100.0	100.0	0.0	14.0	11.9	-	-
	ISIS-ALASSO	100.0	100.0	0.0	17.0	11.7	-	-
	OMP	100.0	99.5	0.0	0.0	33.0	-	-
	S-OMP	100.0	100.0	0.0	100.0	10.0	-	-
	S-OMP-ALASSO	100.0	100.0	0.0	100.0	10.0	-	-
Exact Support	SIS-ALASSO	0.0	100.0	65.5	0.0	2759.0	33.13	0.31
	ISIS-ALASSO	0.0	100.0	62.7	0.0	2984.0	31.71	0.33
	OMP	100.0	100.0	0.0	0.0	8023.1	0.45	0.82
	S-OMP-ALASSO	0.0	100.0	48.1	0.0	4152.9	24.25	0.46
SNR = 1								
Union Support	SIS-ALASSO	0.0	100.0	97.6	0.0	0.2	-	-
	ISIS-ALASSO	0.0	100.0	97.3	0.0	0.3	-	-
	OMP	100.0	99.0	0.0	0.0	59.5	-	-
	S-OMP	0.0	100.0	90.0	0.0	1.0	-	-
	S-OMP-ALASSO	0.0	100.0	93.0	0.0	0.7	-	-
Exact Support	SIS-ALASSO	0.0	100.0	100.0	0.0	0.3	49.94	-0.01
	ISIS-ALASSO	0.0	100.0	100.0	0.0	0.3	49.94	-0.01
	OMP	0.0	100.0	76.4	0.0	1942.8	39.98	0.10
	S-OMP-ALASSO	0.0	100.0	100.0	0.0	1.8	49.94	-0.01

Simulation 2.c: $(n, p, s, T) = (200, 5000, 10, 1000)$, $T_{\text{non-zero}} = 500$

Method name		$\mathcal{M}_* \subseteq \hat{S}$	Correct zeros	Incorrect zeros	$\mathcal{M}_* = \hat{S}$	$ \hat{S} $	$\ \mathbf{B} - \hat{\mathbf{B}}\ _2^2$	R^2
SNR = 10								
Union Support	SIS-ALASSO	100.0	100.0	0.0	89.0	10.1	-	-
	ISIS-ALASSO	100.0	100.0	0.0	95.0	10.1	-	-
	OMP	100.0	99.6	0.0	0.0	29.1	-	-
	S-OMP	100.0	100.0	0.0	100.0	10.0	-	-
	S-OMP-ALASSO	100.0	100.0	0.0	100.0	10.0	-	-
Exact Support	SIS-ALASSO	15.0	100.0	0.2	13.0	4990.2	0.19	0.89
	ISIS-ALASSO	100.0	100.0	0.0	95.0	5000.1	0.12	0.89
	OMP	100.0	100.0	0.0	0.0	5019.2	0.09	0.89
	S-OMP-ALASSO	100.0	100.0	0.0	100.0	5000.0	0.09	0.89
SNR = 5								
Union Support	SIS-ALASSO	100.0	100.0	0.0	27.0	11.4	-	-
	ISIS-ALASSO	100.0	100.0	0.0	14.0	11.6	-	-
	OMP	100.0	99.6	0.0	0.0	29.1	-	-
	S-OMP	100.0	100.0	0.0	100.0	10.0	-	-
	S-OMP-ALASSO	100.0	100.0	0.0	100.0	10.0	-	-
Exact Support	SIS-ALASSO	1.0	100.0	0.8	0.0	4958.9	0.69	0.80
	ISIS-ALASSO	39.0	100.0	0.2	10.0	4991.9	0.44	0.81
	OMP	100.0	100.0	0.0	0.0	5019.2	0.18	0.81
	S-OMP-ALASSO	88.0	100.0	0.0	87.0	4998.8	0.19	0.81
SNR = 1								
Union Support	SIS-ALASSO	0.0	100.0	46.3	0.0	5.4	-	-
	ISIS-ALASSO	1.0	100.0	42.8	1.0	5.7	-	-
	OMP	100.0	99.4	0.0	0.0	38.9	-	-
	S-OMP	0.0	100.0	90.0	0.0	1.0	-	-
	S-OMP-ALASSO	0.0	100.0	90.0	0.0	1.0	-	-
Exact Support	SIS-ALASSO	0.0	100.0	99.8	0.0	8.6	31.16	-0.00
	ISIS-ALASSO	0.0	100.0	99.8	0.0	9.6	31.16	-0.00
	OMP	0.0	100.0	17.5	0.0	4155.6	6.16	0.39
	S-OMP-ALASSO	0.0	100.0	99.6	0.0	20.1	31.11	-0.00

Simulation 2.c: $(n, p, s, T) = (200, 5000, 10, 1000)$, $T_{\text{non-zero}} = 200$

Method name		$\mathcal{M}_* \subseteq \hat{S}$	Correct zeros	Incorrect zeros	$\mathcal{M}_* = \hat{S}$	$ \hat{S} $	$\ \mathbf{B} - \hat{\mathbf{B}}\ _2^2$	R^2
SNR = 10								
Union Support	SIS-ALASSO	100.0	100.0	0.0	100.0	10.0	-	-
	ISIS-ALASSO	100.0	100.0	0.0	100.0	10.0	-	-
	OMP	100.0	97.4	0.0	0.0	139.6	-	-
	S-OMP	100.0	100.0	0.0	100.0	10.0	-	-
	S-OMP-ALASSO	100.0	100.0	0.0	100.0	10.0	-	-
Exact Support	SIS-ALASSO	100.0	100.0	0.0	100.0	2000.0	0.02	0.79
	ISIS-ALASSO	100.0	100.0	0.0	100.0	2000.0	0.02	0.79
	OMP	100.0	100.0	0.0	0.0	2131.6	0.03	0.78
	S-OMP-ALASSO	100.0	100.0	0.0	100.0	2000.0	0.01	0.79
SNR = 5								
Union Support	SIS-ALASSO	100.0	100.0	0.0	100.0	10.0	-	-
	ISIS-ALASSO	100.0	100.0	0.0	100.0	10.0	-	-
	OMP	100.0	97.4	0.0	0.0	139.6	-	-
	S-OMP	100.0	100.0	0.0	100.0	10.0	-	-
	S-OMP-ALASSO	100.0	100.0	0.0	100.0	10.0	-	-
Exact Support	SIS-ALASSO	100.0	100.0	0.0	100.0	2000.0	0.04	0.72
	ISIS-ALASSO	100.0	100.0	0.0	100.0	2000.0	0.04	0.72
	OMP	100.0	100.0	0.0	0.0	2131.6	0.05	0.71
	S-OMP-ALASSO	100.0	100.0	0.0	100.0	2000.0	0.03	0.72
SNR = 1								
Union Support	SIS-ALASSO	100.0	100.0	0.0	37.0	11.1	-	-
	ISIS-ALASSO	100.0	100.0	0.0	44.0	10.8	-	-
	OMP	100.0	97.4	0.0	0.0	139.6	-	-
	S-OMP	0.0	100.0	90.0	0.0	1.0	-	-
	S-OMP-ALASSO	0.0	100.0	90.0	0.0	1.0	-	-
Exact Support	SIS-ALASSO	0.0	100.0	12.0	0.0	1761.3	2.15	0.37
	ISIS-ALASSO	0.0	100.0	9.1	0.0	1819.3	1.71	0.38
	OMP	99.0	100.0	0.0	0.0	2131.6	0.26	0.42
	S-OMP-ALASSO	0.0	100.0	93.2	0.0	136.0	11.65	0.03

Simulation 3: $(n, p, s, T) = (100, 5000, 3, 150)$, $T_{\text{non-zero}} = 80$, $\rho = 0.2$

Method name		$\mathcal{M}_* \subseteq \hat{S}$	Correct zeros	Incorrect zeros	$\mathcal{M}_* = \hat{S}$	$ \hat{S} $	$\ \mathbf{B} - \hat{\mathbf{B}}\ _2^2$	R^2
SNR = 10								
Union Support	SIS-ALASSO	100.0	100.0	0.0	100.0	3.0	-	-
	ISIS-ALASSO	100.0	100.0	0.0	100.0	3.0	-	-
	OMP	100.0	99.8	0.0	0.0	20.0	-	-
	S-OMP	100.0	100.0	0.0	100.0	3.0	-	-
	S-OMP-ALASSO	100.0	100.0	0.0	100.0	3.0	-	-
Exact Support	SIS-ALASSO	96.0	100.0	0.0	96.0	239.9	0.02	0.73
	ISIS-ALASSO	100.0	100.0	0.0	100.0	240.0	0.02	0.73
	OMP	100.0	100.0	0.0	0.0	257.1	0.03	0.72
	S-OMP-ALASSO	100.0	100.0	0.0	100.0	240.0	0.01	0.73
SNR = 5								
Union Support	SIS-ALASSO	100.0	100.0	0.0	100.0	3.0	-	-
	ISIS-ALASSO	100.0	100.0	0.0	100.0	3.0	-	-
	OMP	100.0	99.8	0.0	0.0	19.6	-	-
	S-OMP	100.0	100.0	0.0	100.0	3.0	-	-
	S-OMP-ALASSO	100.0	100.0	0.0	100.0	3.0	-	-
Exact Support	SIS-ALASSO	100.0	100.0	0.0	100.0	240.0	0.02	0.72
	ISIS-ALASSO	100.0	100.0	0.0	100.0	240.0	0.02	0.72
	OMP	100.0	100.0	0.0	0.0	256.6	0.03	0.72
	S-OMP-ALASSO	100.0	100.0	0.0	100.0	240.0	0.01	0.72
SNR = 1								
Union Support	SIS-ALASSO	100.0	100.0	0.0	92.0	3.1	-	-
	ISIS-ALASSO	100.0	100.0	0.0	94.0	3.1	-	-
	OMP	100.0	99.8	0.0	0.0	20.3	-	-
	S-OMP	100.0	100.0	0.0	100.0	3.0	-	-
	S-OMP-ALASSO	100.0	100.0	0.0	100.0	3.0	-	-
Exact Support	SIS-ALASSO	99.0	100.0	0.0	92.0	240.1	0.04	0.70
	ISIS-ALASSO	100.0	100.0	0.0	94.0	240.1	0.03	0.70
	OMP	100.0	100.0	0.0	0.0	257.3	0.04	0.69
	S-OMP-ALASSO	100.0	100.0	0.0	100.0	240.0	0.02	0.70

Simulation 3: $(n, p, s, T) = (100, 5000, 3, 150)$, $T_{\text{non-zero}} = 80$, $\rho = 0.5$

Method name		$\mathcal{M}_* \subseteq \hat{S}$	Correct zeros	Incorrect zeros	$\mathcal{M}_* = \hat{S}$	$ \hat{S} $	$\ \mathbf{B} - \hat{\mathbf{B}}\ _2^2$	R^2
SNR = 10								
Union Support	SIS-ALASSO	100.0	100.0	0.0	98.0	3.0	-	-
	ISIS-ALASSO	100.0	100.0	0.0	100.0	3.0	-	-
	OMP	100.0	99.8	0.0	0.0	20.1	-	-
	S-OMP	100.0	100.0	0.0	100.0	3.0	-	-
	S-OMP-ALASSO	100.0	100.0	0.0	100.0	3.0	-	-
Exact Support	SIS-ALASSO	87.0	100.0	0.2	85.0	239.5	0.08	0.62
	ISIS-ALASSO	88.0	100.0	0.1	88.0	239.8	0.07	0.62
	OMP	100.0	100.0	0.0	0.0	257.1	0.06	0.62
	S-OMP-ALASSO	100.0	100.0	0.0	100.0	240.0	0.03	0.63
SNR = 5								
Union Support	SIS-ALASSO	100.0	100.0	0.0	97.0	3.0	-	-
	ISIS-ALASSO	100.0	100.0	0.0	96.0	3.0	-	-
	OMP	100.0	99.8	0.0	0.0	19.6	-	-
	S-OMP	100.0	100.0	0.0	100.0	3.0	-	-
	S-OMP-ALASSO	100.0	100.0	0.0	100.0	3.0	-	-
Exact Support	SIS-ALASSO	60.0	100.0	0.2	57.0	239.5	0.10	0.61
	ISIS-ALASSO	84.0	100.0	0.1	80.0	239.8	0.08	0.61
	OMP	100.0	100.0	0.0	0.0	256.6	0.06	0.61
	S-OMP-ALASSO	100.0	100.0	0.0	100.0	240.0	0.03	0.62
SNR = 1								
Union Support	SIS-ALASSO	100.0	100.0	0.0	56.0	3.5	-	-
	ISIS-ALASSO	100.0	100.0	0.0	70.0	3.4	-	-
	OMP	100.0	99.8	0.0	0.0	19.9	-	-
	S-OMP	100.0	100.0	0.0	100.0	3.0	-	-
	S-OMP-ALASSO	100.0	100.0	0.0	100.0	3.0	-	-
Exact Support	SIS-ALASSO	1.0	100.0	2.3	1.0	235.1	0.21	0.58
	ISIS-ALASSO	5.0	100.0	1.5	3.0	236.8	0.16	0.58
	OMP	96.0	100.0	0.0	0.0	256.9	0.08	0.58
	S-OMP-ALASSO	67.0	100.0	0.2	67.0	239.5	0.05	0.59

Simulation 3: $(n, p, s, T) = (100, 5000, 3, 150)$, $T_{\text{non-zero}} = 80$, $\rho = 0.7$

Method name		$\mathcal{M}_* \subseteq \hat{S}$	Correct zeros	Incorrect zeros	$\mathcal{M}_* = \hat{S}$	$ \hat{S} $	$\ \mathbf{B} - \hat{\mathbf{B}}\ _2^2$	R^2
SNR = 10								
Union Support	SIS-ALASSO	80.0	100.0	6.7	80.0	2.8	-	-
	ISIS-ALASSO	85.0	100.0	5.0	85.0	2.9	-	-
	OMP	100.0	99.8	0.0	0.0	22.0	-	-
	S-OMP	0.0	100.0	51.0	0.0	1.5	-	-
	S-OMP-ALASSO	0.0	100.0	51.0	0.0	1.5	-	-
Exact Support	SIS-ALASSO	0.0	100.0	63.3	0.0	88.1	3.93	0.15
	ISIS-ALASSO	0.0	100.0	61.0	0.0	93.6	3.70	0.16
	OMP	0.0	100.0	12.0	0.0	230.2	0.73	0.28
	S-OMP-ALASSO	0.0	100.0	57.6	0.0	101.8	2.89	0.19
SNR = 5								
Union Support	SIS-ALASSO	79.0	100.0	7.0	79.0	2.8	-	-
	ISIS-ALASSO	85.0	100.0	5.0	83.0	2.9	-	-
	OMP	100.0	99.8	0.0	0.0	22.5	-	-
	S-OMP	0.0	100.0	56.7	0.0	1.3	-	-
	S-OMP-ALASSO	0.0	100.0	56.7	0.0	1.3	-	-
Exact Support	SIS-ALASSO	0.0	100.0	66.0	0.0	81.6	4.15	0.14
	ISIS-ALASSO	0.0	100.0	64.2	0.0	85.9	3.95	0.15
	OMP	0.0	100.0	16.5	0.0	219.8	0.96	0.26
	S-OMP-ALASSO	0.0	100.0	61.2	0.0	93.0	3.16	0.18
SNR = 1								
Union Support	SIS-ALASSO	89.0	100.0	3.7	45.0	3.5	-	-
	ISIS-ALASSO	92.0	100.0	2.7	49.0	3.5	-	-
	OMP	100.0	99.8	0.0	0.0	27.7	-	-
	S-OMP	0.0	100.0	60.3	0.0	1.2	-	-
	S-OMP-ALASSO	0.0	100.0	60.3	0.0	1.2	-	-
Exact Support	SIS-ALASSO	0.0	100.0	71.4	0.0	69.4	4.76	0.11
	ISIS-ALASSO	0.0	100.0	68.9	0.0	75.3	4.46	0.12
	OMP	0.0	100.0	29.3	0.0	196.8	1.96	0.23
	S-OMP-ALASSO	0.0	100.0	64.6	0.0	85.0	3.53	0.16

Simulation 4: $(n, p, s, T) = (150, 4000, 8, 150)$, $T_{\text{non-zero}} = 80$, $\rho = 0.2$

Method name		$\mathcal{M}_* \subseteq \hat{S}$	Correct zeros	Incorrect zeros	$\mathcal{M}_* = \hat{S}$	$ \hat{S} $	$\ \mathbf{B} - \hat{\mathbf{B}}\ _2^2$	R^2
SNR = 10								
Union Support	SIS-ALASSO	100.0	100.0	0.0	100.0	8.0	-	-
	ISIS-ALASSO	100.0	100.0	0.0	97.0	8.0	-	-
	OMP	100.0	99.9	0.0	2.0	11.7	-	-
	S-OMP	100.0	100.0	0.0	100.0	8.0	-	-
	S-OMP-ALASSO	100.0	100.0	0.0	100.0	8.0	-	-
Exact Support	SIS-ALASSO	35.0	100.0	1.4	35.0	631.3	0.55	0.88
	ISIS-ALASSO	100.0	100.0	0.0	97.0	640.0	0.14	0.89
	OMP	100.0	100.0	0.0	2.0	643.7	0.10	0.89
	S-OMP-ALASSO	100.0	100.0	0.0	100.0	640.0	0.09	0.89
SNR = 5								
Union Support	SIS-ALASSO	100.0	100.0	0.0	85.0	8.2	-	-
	ISIS-ALASSO	100.0	100.0	0.0	78.0	8.3	-	-
	OMP	100.0	99.9	0.0	2.0	11.7	-	-
	S-OMP	100.0	100.0	0.0	100.0	8.0	-	-
	S-OMP-ALASSO	100.0	100.0	0.0	100.0	8.0	-	-
Exact Support	SIS-ALASSO	2.0	100.0	4.5	2.0	611.7	1.78	0.77
	ISIS-ALASSO	7.0	100.0	2.9	6.0	621.5	1.29	0.78
	OMP	100.0	100.0	0.0	2.0	643.7	0.20	0.80
	S-OMP-ALASSO	39.0	100.0	1.0	39.0	633.8	0.48	0.80
SNR = 1								
Union Support	SIS-ALASSO	0.0	100.0	90.5	0.0	0.8	-	-
	ISIS-ALASSO	0.0	100.0	87.6	0.0	1.0	-	-
	OMP	100.0	99.8	0.0	0.0	14.9	-	-
	S-OMP	0.0	100.0	87.5	0.0	1.0	-	-
	S-OMP-ALASSO	0.0	100.0	88.5	0.0	0.9	-	-
Exact Support	SIS-ALASSO	0.0	100.0	99.9	0.0	0.8	29.62	-0.01
	ISIS-ALASSO	0.0	100.0	99.8	0.0	1.1	29.61	-0.01
	OMP	0.0	100.0	31.1	0.0	447.7	10.11	0.32
	S-OMP-ALASSO	0.0	100.0	99.6	0.0	2.7	29.56	-0.00

Simulation 4: $(n, p, s, T) = (150, 4000, 8, 150)$, $T_{\text{non-zero}} = 80$, $\rho = 0.5$

Method name		$\mathcal{M}_* \subseteq \hat{S}$	Correct zeros	Incorrect zeros	$\mathcal{M}_* = \hat{S}$	$ \hat{S} $	$\ \mathbf{B} - \hat{\mathbf{B}}\ _2^2$	R^2
SNR = 10								
Union Support	SIS-ALASSO	100.0	100.0	0.0	80.0	8.2	-	-
	ISIS-ALASSO	100.0	100.0	0.0	89.0	8.1	-	-
	OMP	100.0	99.9	0.0	2.0	11.9	-	-
	S-OMP	100.0	100.0	0.0	100.0	8.0	-	-
	S-OMP-ALASSO	100.0	100.0	0.0	100.0	8.0	-	-
Exact Support	SIS-ALASSO	0.0	100.0	13.1	0.0	556.5	4.24	0.80
	ISIS-ALASSO	80.0	100.0	0.2	70.0	638.9	0.23	0.89
	OMP	100.0	100.0	0.0	2.0	643.9	0.11	0.89
	S-OMP-ALASSO	100.0	100.0	0.0	100.0	640.0	0.10	0.89
SNR = 5								
Union Support	SIS-ALASSO	100.0	100.0	0.0	69.0	8.4	-	-
	ISIS-ALASSO	100.0	100.0	0.0	47.0	8.9	-	-
	OMP	100.0	99.9	0.0	2.0	12.3	-	-
	S-OMP	100.0	100.0	0.0	100.0	8.0	-	-
	S-OMP-ALASSO	100.0	100.0	0.0	100.0	8.0	-	-
Exact Support	SIS-ALASSO	0.0	100.0	23.8	0.0	487.8	7.53	0.65
	ISIS-ALASSO	0.0	100.0	7.6	0.0	592.5	2.75	0.75
	OMP	99.0	100.0	0.0	2.0	644.4	0.22	0.80
	S-OMP-ALASSO	7.0	100.0	2.8	7.0	622.2	1.04	0.79
SNR = 1								
Union Support	SIS-ALASSO	0.0	100.0	60.6	0.0	3.2	-	-
	ISIS-ALASSO	1.0	100.0	56.8	1.0	3.5	-	-
	OMP	100.0	99.6	0.0	0.0	23.5	-	-
	S-OMP	0.0	100.0	87.5	0.0	1.0	-	-
	S-OMP-ALASSO	0.0	100.0	87.5	0.0	1.0	-	-
Exact Support	SIS-ALASSO	0.0	100.0	99.3	0.0	4.7	29.45	-0.00
	ISIS-ALASSO	0.0	100.0	99.2	0.0	5.1	29.43	-0.00
	OMP	0.0	100.0	44.9	0.0	369.3	15.05	0.28
	S-OMP-ALASSO	0.0	100.0	98.5	0.0	9.9	29.39	0.01

Simulation 5: $(n, p, s, T) = (200, 10000, 5, 500)$, $T_{\text{non-zero}} = 400$

Method name		$\mathcal{M}_* \subseteq \hat{S}$	Correct zeros	Incorrect zeros	$\mathcal{M}_* = \hat{S}$	$ \hat{S} $	$\ \mathbf{B} - \hat{\mathbf{B}}\ _2^2$	R^2
$\sigma = 1.5$								
Union Support	SIS-ALASSO	53.0	99.6	9.4	0.0	41.1	-	-
	ISIS-ALASSO	100.0	99.8	0.0	0.0	28.1	-	-
	OMP	100.0	99.9	0.0	12.0	10.0	-	-
	S-OMP	100.0	100.0	0.0	44.0	5.6	-	-
	S-OMP-ALASSO	100.0	100.0	0.0	100.0	5.0	-	-
Exact Support	SIS-ALASSO	0.0	100.0	68.9	0.0	936.0	84.66	0.66
	ISIS-ALASSO	0.0	100.0	16.2	0.0	1791.9	5.80	0.96
	OMP	100.0	100.0	0.0	12.0	2090.3	0.06	0.99
	S-OMP-ALASSO	100.0	100.0	0.0	100.0	2000.0	0.05	0.99
$\sigma = 2.5$								
Union Support	SIS-ALASSO	53.0	99.4	9.4	0.0	61.4	-	-
	ISIS-ALASSO	100.0	99.3	0.0	0.0	77.7	-	-
	OMP	100.0	99.9	0.0	10.0	13.2	-	-
	S-OMP	100.0	100.0	0.0	44.0	5.6	-	-
	S-OMP-ALASSO	100.0	100.0	0.0	100.0	5.0	-	-
Exact Support	SIS-ALASSO	0.0	100.0	69.2	0.0	910.2	85.82	0.64
	ISIS-ALASSO	0.0	100.0	17.5	0.0	1834.1	7.23	0.93
	OMP	100.0	100.0	0.0	10.0	2093.3	0.16	0.96
	S-OMP-ALASSO	93.0	100.0	0.0	93.0	1999.9	0.13	0.96
$\sigma = 4.5$								
Union Support	SIS-ALASSO	40.0	99.1	12.0	0.0	92.5	-	-
	ISIS-ALASSO	100.0	97.8	0.0	0.0	226.8	-	-
	OMP	100.0	99.8	0.0	1.0	25.7	-	-
	S-OMP	92.0	100.0	1.6	46.0	5.5	-	-
	S-OMP-ALASSO	92.0	100.0	1.6	92.0	5.0	-	-
Exact Support	SIS-ALASSO	0.0	100.0	70.0	0.0	850.2	88.65	0.56
	ISIS-ALASSO	0.0	100.0	27.4	0.0	1847.2	15.79	0.83
	OMP	0.0	100.0	3.2	0.0	2040.9	1.15	0.88
	S-OMP-ALASSO	0.0	100.0	10.2	0.0	1795.3	2.38	0.87

Simulation 5: $(n, p, s, T) = (200, 10000, 5, 500)$, $T_{\text{non-zero}} = 250$

Method name		$\mathcal{M}_* \subseteq \hat{S}$	Correct zeros	Incorrect zeros	$\mathcal{M}_* = \hat{S}$	$ \hat{S} $	$\ \mathbf{B} - \hat{\mathbf{B}}\ _2^2$	R^2
$\sigma = 1.5$								
Union Support	SIS-ALASSO	100.0	99.7	0.0	0.0	31.5	-	-
	ISIS-ALASSO	100.0	99.9	0.0	1.0	14.3	-	-
	OMP	100.0	99.7	0.0	0.0	30.8	-	-
	S-OMP	100.0	100.0	0.0	20.0	5.8	-	-
	S-OMP-ALASSO	100.0	100.0	0.0	100.0	5.0	-	-
Exact Support	SIS-ALASSO	0.0	100.0	45.9	0.0	768.9	25.98	0.79
	ISIS-ALASSO	0.0	100.0	5.3	0.0	1200.7	1.00	0.92
	OMP	100.0	100.0	0.0	0.0	1287.6	0.05	0.92
	S-OMP-ALASSO	100.0	100.0	0.0	100.0	1250.0	0.03	0.92
$\sigma = 2.5$								
Union Support	SIS-ALASSO	100.0	99.6	0.0	0.0	40.5	-	-
	ISIS-ALASSO	100.0	99.6	0.0	0.0	44.3	-	-
	OMP	100.0	99.7	0.0	0.0	32.0	-	-
	S-OMP	100.0	100.0	0.0	23.0	5.8	-	-
	S-OMP-ALASSO	100.0	100.0	0.0	100.0	5.0	-	-
Exact Support	SIS-ALASSO	0.0	100.0	46.2	0.0	757.5	26.30	0.74
	ISIS-ALASSO	0.0	100.0	7.5	0.0	1205.2	1.55	0.86
	OMP	100.0	100.0	0.0	0.0	1288.6	0.14	0.87
	S-OMP-ALASSO	92.0	100.0	0.0	92.0	1249.9	0.08	0.87
$\sigma = 4.5$								
Union Support	SIS-ALASSO	98.0	99.6	0.4	0.0	48.0	-	-
	ISIS-ALASSO	100.0	99.0	0.0	0.0	104.0	-	-
	OMP	100.0	99.7	0.0	0.0	36.1	-	-
	S-OMP	1.0	100.0	19.8	1.0	4.7	-	-
	S-OMP-ALASSO	1.0	100.0	19.8	1.0	4.2	-	-
Exact Support	SIS-ALASSO	0.0	100.0	48.4	0.0	713.1	27.64	0.62
	ISIS-ALASSO	0.0	100.0	22.8	0.0	1080.7	5.57	0.71
	OMP	0.0	100.0	2.3	0.0	1264.0	0.70	0.75
	S-OMP-ALASSO	0.0	100.0	19.9	0.0	1002.0	2.26	0.73

Simulation 5: $(n, p, s, T) = (200, 10000, 5, 500)$, $T_{\text{non-zero}} = 100$

Method name		$\mathcal{M}_* \subseteq \hat{S}$	Correct zeros	Incorrect zeros	$\mathcal{M}_* = \hat{S}$	$ \hat{S} $	$\ \mathbf{B} - \hat{\mathbf{B}}\ _2^2$	R^2
$\sigma = 1.5$								
Union Support	SIS-ALASSO	100.0	99.9	0.0	1.0	10.9	-	-
	ISIS-ALASSO	100.0	100.0	0.0	56.0	5.7	-	-
	OMP	100.0	98.0	0.0	0.0	205.8	-	-
	S-OMP	99.0	100.0	0.2	4.0	6.0	-	-
	S-OMP-ALASSO	99.0	100.0	0.2	99.0	5.0	-	-
Exact Support	SIS-ALASSO	0.0	100.0	19.4	0.0	411.0	2.86	0.60
	ISIS-ALASSO	17.0	100.0	0.5	16.0	498.0	0.06	0.62
	OMP	100.0	100.0	0.0	0.0	726.4	0.19	0.60
	S-OMP-ALASSO	99.0	100.0	0.2	99.0	499.0	0.02	0.62
$\sigma = 2.5$								
Union Support	SIS-ALASSO	100.0	99.9	0.0	1.0	11.0	-	-
	ISIS-ALASSO	100.0	99.9	0.0	0.0	12.4	-	-
	OMP	100.0	98.0	0.0	0.0	205.8	-	-
	S-OMP	0.0	100.0	20.0	0.0	4.9	-	-
	S-OMP-ALASSO	0.0	100.0	20.0	0.0	4.0	-	-
Exact Support	SIS-ALASSO	0.0	100.0	19.6	0.0	408.8	2.92	0.54
	ISIS-ALASSO	0.0	100.0	2.5	0.0	495.2	0.21	0.56
	OMP	100.0	100.0	0.0	0.0	726.4	0.54	0.53
	S-OMP-ALASSO	0.0	100.0	20.0	0.0	400.0	0.83	0.52
$\sigma = 4.5$								
Union Support	SIS-ALASSO	98.0	100.0	0.4	1.0	9.8	-	-
	ISIS-ALASSO	97.0	99.9	0.6	0.0	17.4	-	-
	OMP	100.0	98.0	0.0	0.0	206.4	-	-
	S-OMP	0.0	100.0	41.2	0.0	3.6	-	-
	S-OMP-ALASSO	0.0	100.0	41.2	0.0	3.4	-	-
Exact Support	SIS-ALASSO	0.0	100.0	27.6	0.0	367.3	3.48	0.41
	ISIS-ALASSO	0.0	100.0	19.9	0.0	413.1	1.33	0.42
	OMP	4.0	100.0	1.4	0.0	720.0	1.79	0.41
	S-OMP-ALASSO	0.0	100.0	41.2	0.0	295.9	4.66	0.35

References

- Andreas Argyriou, Theodoros Evgeniou, and Massimiliano Pontil. Convex multi-task feature learning. *Machine Learning*, 73(3): 243–272, December 2008. doi: 10.1007/s10994-007-5040-8. URL <http://dx.doi.org/10.1007/s10994-007-5040-8>.
- Andrew R. Barron, Albert Cohen, Wolfgang Dahmen, and Ronald A. DeVore. Approximation and learning by greedy algorithms. *The Annals of Statistics*, 36(1):64–94, 2008. doi: 10.1214/009053607000000631. URL <http://projecteuclid.org/DPubS?service=UI&version=1.0&verb=Display&handle=euclid.aos/12018>
- Peter J. Bickel, Ya’acov Ritov, and Alexandre B. Tsybakov. Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics*, 37(4):1705–1732, 2009. doi: 10.1214/08-AOS620. URL <http://projecteuclid.org/DPubS?service=UI&version=1.0&verb=Display&handle=euclid.aos/12453>
- Jiahua Chen and Zehua Chen. Extended Bayesian information criteria for model selection with large model spaces. *Biometrika*, 95(3):759–771, 2008. doi: 10.1093/biomet/asn034. URL <http://biomet.oxfordjournals.org/cgi/content/abstract/95/3/759>.
- S.F. Cotter, R. Adler, R.D. Rao, and K. Kreutz-Delgado. Forward sequential algorithms for best basis selection. *Vision, Image and Signal Processing, IEE Proceedings* -, 146(5):235–244, 1999. ISSN 1350-245X. doi: 10.1049/ip-vis:19990445.
- Jianqing Fan and Jinchi Lv. Sure independence screening for ultrahigh dimensional feature space. *Journal Of The Royal Statistical Society Series B*, 70(5):849–911, 2008. URL <http://ideas.repec.org/a/bla/jorssb/v70y2008i5p849-911.html>.
- J. Huang, S. Ma, and C. H Zhang. Adaptive lasso for sparse High-Dimensional regression models. *Statistica Sinica*, 18:16031618, 2008.
- Seyoung Kim and Eric P. Xing. Statistical estimation of correlated genome associations to a quantitative trait network. *PLoS Genet*, 5(8):e1000587, 2009. doi: 10.1371/journal.pgen.1000587. URL <http://dx.doi.org/10.1371/journal.pgen.1000587>.
- Seyoung Kim, Kyung-Ah Sohn, and Eric P. Xing. A multivariate regression approach to association analysis of a quantitative trait network. *Bioinformatics*, 25(12):i204–212, June 2009. doi: 10.1093/bioinformatics/btp218. URL <http://bioinformatics.oxfordjournals.org/cgi/content/abstract/25/12/i204>.
- M. Kolar, J. Lafferty, and L. Wasserman. Union Support Recovery in Multi-task Learning. *ArXiv e-prints*, August 2010.

Béatrice Laurent and Pascal Massart. Adaptive estimation of a quadratic functional by model selection. *Annals of Statistics*, 28(5):1302–1338, 2000. doi: [doi:10.1214/aos/1015957395](https://doi.org/10.1214/aos/1015957395).

Han Liu, Mark Palatucci, and Jian Zhang. Blockwise coordinate descent procedures for the multi-task lasso, with applications to neural semantic basis discovery. In *ICML '09: Proceedings of the 26th Annual International Conference on Machine Learning*, pages 649–656, New York, NY, USA, 2009. ACM. ISBN 978-1-60558-516-1. doi: <http://doi.acm.org/10.1145/1553374.1553458>.

Karim Lounici, Massimiliano Pontil, Alexandre B. Tsybakov, and Sara van de Geer. Taking advantage of sparsity in Multi-Task learning. In *Proceedings of the Conference on Learning Theory (COLT)*, 2009. URL <http://arxiv.org/abs/0903.1468>.

A. Lozano, G. Swirszcz, and N. Abe. Grouped orthogonal matching pursuit for variable selection and prediction. In *Advances in Neural Information Processing Systems 22*. 2009.

Sahand Negahban and Martin Wainwright. Phase transitions for high-dimensional joint support recovery. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems 21*, pages 1161–1168. 2009.

G. Obozinski, M.J. Wainwright, and M.I. Jordan. Support union recovery in high-dimensional multivariate regression. *Annals of Statistics*, to appear, 2010.

Guillaume Obozinski, Martin Wainwright, and Michael Jordan. High-dimensional support union recovery in multivariate regression. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems 21*, pages 1217–1224. 2009.

Jie Peng, Ji Zhu, Anna Bergamaschi, Wonshik Han, Dong-Young Noh, Jonathan R. Pollack, and Pei Wang. Regularized multivariate regression for identifying master predictors with application to integrative genomics study of breast cancer, 2008. URL <http://www.citebase.org/abstract?id=oai:arXiv.org:0812.3671>.

R. Rubinstein, M. Zibulevsky, and M. Elad. Efficient implementation of the K-SVD algorithm using batch orthogonal matching pursuit. *Technical report*, 2008.

Joel A. Tropp, Anna C. Gilbert, and Martin J. Strauss. Algorithms for simultaneous sparse approximation. part i: Greedy pursuit. *Signal Processing*, 86(3): 572 – 588, 2006. ISSN 0165-1684. doi: DOI:10.1016/j.sigpro.2005.05.030. URL <http://www.sciencedirect.com/science/article/B6V18-4GWC8JH-1/2/356d996af09dd87d495a94fba1>. Sparse Approximations in Signal and Image Processing.

- Hansheng Wang. Forward Regression for Ultra-High Dimensional Variable Screening. *SSRN eLibrary*, 2009.
- H Zha, C Ding, M Gu, X He, and H Simon. Spectral relaxation for k-means clustering. pages 1057–1064. MIT Press, 2001.
- Cun-Hui Zhang and Jian Huang. The sparsity and bias of the lasso selection in high-dimensional linear regression. *Annals of Statistics*, 36(4):1567–1594, 2008. URL doi:10.1214/07-AOS520.
- J. Zhang. *A probabilistic framework for multitask learning (Technical Report CMU-LTI-06-006)*. PhD thesis, Ph. D. thesis, Carnegie Mellon University, 2006.
- Tong Zhang. On the consistency of feature selection using greedy least squares regression. *J. Mach. Learn. Res.*, 10(Mar):555–568, 2009.
- Shuheng Zhou, Sara van de Geer, and Peter Buhlmann. Adaptive lasso for high dimensional regression and gaussian graphical modeling, 2009. URL <http://www.citebase.org/abstract?id=oai:arXiv.org:0903.2515>.
- Hui Zou. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101:1418–1429, December 2006. URL <http://ideas.repec.org/a/bes/jnlasa/v101y2006p1418-1429.html>.